Department of History and Philosophy of Science

University of Cambridge

# *Representing and Constructing.*

## *Psychometrics from the perspectives of measurement theory and concept formation.*

This dissertation is submitted for the degree of Doctor of Philosophy.

Elina Sini Maria Vessonen

Newnham College

September 2018

Dissertation title: *Representing and Constructing. Psychometrics from the perspectives of measurement theory and concept formation.*

Author: Elina Sini Maria Vessonen

## Abstract

Social scientific measurement is much desired and much criticized. In this dissertation I evaluate one of the main approaches to social scientific measurement that has nevertheless been virtually ignored by philosophers – the psychometric approach. Psychometric measures are typically used to measure unobservable attributes such as intelligence and personality. They typically take the form of questionnaires or tests and are validated by statistical tests of properties such as reliability and model-fit.

My thesis is two-fold. In the first, more critical part, I argue that psychometric instruments normally fail to fulfil plausible criteria for adequate measurement. To make this argument, I define and defend a conception of quantitative representation necessary for measurement. My definition is grounded in the Representational Theory of Measurement but avoids the main critiques this theory has faced. I then show that the typical psychometric process of measure validation fails to produce evidence of such quantitative representation. The upshot is that although a quantitative interpretation of psychometric data is common, it is largely unwarranted.

In the second part, I argue that psychometric instruments are nonetheless apt for various other purposes. This argument hinges on a new outlook on how concepts should be formed for psychometric purposes. Philosophers have traditionally thought that concepts should cohere with intuitions and/or pick out so-called natural kinds, while many psychometricians argue that concepts should pick out real as opposed to constructed attributes. I argue that, when it comes to social scientific measurement, it is much more important to focus on the usefulness of the concept, where usefulness can take on different meanings in different contexts. Building on the defended outlook on concept formation, I show what useful functions psychometric instruments can serve even when they fail at quantitative representation.

# Declaration of length

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Introduction and specified in the text.

This dissertation is not substantially the same as any that I have submitted, or, that is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Introduction and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Introduction and specified in the text.

This dissertation does not exceed the prescribed word limit of 80,000 words, including footnotes but excluding bibliography.

# Acknowledgments

I applied to Cambridge to learn from Anna Alexandrova and Hasok Chang, who kindly agreed to be my PhD supervisors. Anna has taught me everything I know about philosophical writing that is concrete, exciting and connected to real life. Our conversations about philosophy of social sciences, as well as the professional opportunities she pointed me to, have enriched my thinking. Hasok – whose book *Inventing Temperature* was what first sparked my interest in philosophy of measurement – has taught me immensely about measurement and philosophy of science in practice. His questions were always difficult and to the point, which challenged me to formulate better arguments. I thank Anna and Hasok for all their help.

Lukas Beck, Juha Haaja, Mark Pender-Bare, Kristiina Tammisalo and Osmo Tammisalo read and commented on earlier drafts of this dissertation. I thank them for their valuable help.

I am grateful to Denny Borsboom and Tim Lewens who provided helpful comments in their capacity as examiners of the thesis. Their questions and criticisms will help me develop further the ideas expressed in this dissertation.

Over the past three years, various people have given feedback on my draft papers and conference presentations, thereby helping me develop the ideas I present in this dissertation. Conversations with academics and my graduate student peers have also taught me a lot and helped me formulate my arguments. I thank especially Erik Angner, Georg Brun, Christopher Clarke, Richard Creath, Conrad Heilmann and Eran Tal as well as Sam Bagg, Lukas Beck, Fons Dewulf, Mats Ingelström, Tanya Paes and Karoliina Pulkkinen for their help.

I thank the following institutions for funding my PhD research: Arts and Humanities Research Council, the British Society for the Philosophy of Science, Cambridge Trust and Newnham College.

I am grateful to my family for supporting me more than with mere words. I thank my friends for the welcome distraction from contemplation. Finally, I am thankful to Juha, who helps me find the right *beta*.

# List of abbreviations

2PL – Two-parameter model in IRT

APA – American Psychological Association

CTT – Classical Test Theory

DAB – Direct Axiom-Based Approach

HAM-D – Hamilton Depression Rating Scale

IMM – Indirect Measurement Model Based Approach

IRC – Item Response Curve

IRT – Item Response Theory

PROM – Patient Rated Outcome Measure

ReMi – Representation Minimalism

RO – Respectful Operationalism

RTM – Representational Theory of Measurement

SWLS – Satisfaction with Life Scale

# List of tables

# List of figures

# Contents

# 1. Introduction

## 1.1 Much desired and much despised

Most people do not know what the term 'psychometrics' denotes. Nonetheless, most of us live lives that are significantly shaped by psychometric measures. Millions of people take anti-depressants, which have been allowed on the market partly based on psychometric measurement of depression (e.g. Keller et al. 2001; Le Noury et al. 2015). Many government institutions, such as the UK Treasury, use psychometric measures of well-being and life satisfaction to inform their policy decisions (e.g. Fujiwara and Campbell 2011). Job applicants are frequently assessed using psychometric instruments – such as the infamous *Myers-Briggs Type Indicator* or the *Minnesota Multiphasic Personality Inventory* – and their suitability for the job determined partly based on the results (e.g. Emre 2018). Most recently, psychometric instruments have found a defining role in mass communication: when Facebook likes are correlated with psychometric tests of personality, the result is a tool for targeting and influencing millions of people based on their psychological profiles (e.g. Matz et al. 2017). The list could go on. Evidently, decision-makers of various dispositions desire psychometric tools.

The historical roots of psychometrics are also entangled with the needs of decision-makers. Although it is technically possible to regard the founding of *the Psychometric Society* (in 1935 in Michigan) as the dawn of psychometrics, the approach grew from earlier developments in the testing of psychological attributes (Jones and Thissen 2006). When the U.S. entered the First World War, the government was in dire need of large-scale assessment of army recruits and their proneness to psychological problems. Among the researchers developing these tests were Edward L. Thorndike and L. L. Thurstone, both of whom later became presidents of the Psychometric Society, and grand names in psychometrics more generally. Before the formal founding of the Psychometric Society, Thurstone also helped develop tests for the selection of the U.S. government personnel as well as examinations of the abilities of high school graduates. While working on these tools for supporting practical decision-making, he simultaneously developed techniques for validating and interpreting psychological tests, such as multiple

factor analysis. The history of psychometrics is littered with similar examples, where the development of tests and techniques for validating them, go hand in hand with the needs of decision-makers.

Psychometric tools have been, and still are, in great demand. But even the most popular psychometric tests are heavily criticized, as if to balance all the hype around them. For example, in their often-cited review article, Bagby et al. (2004) amass an impressive range of evidence *against* the validity of the *Hamilton Depression Rating Scale* (HAM-D), a measure that is used in the vast majority of anti-depressant trials. The wildly popular Myers-Briggs test – a brainchild of a mother-daughter-pair with no formal education in psychology but a deep fascination with Carl Jung's work – is heavily criticized as invalid and unreliable in the academic literature (e.g. Pittenger 2005; see Emre 2018 on the history of the test).[1] Reports of national happiness, too, tend to be met with scepticism. The results of *the World Happiness Report*, which ranks countries according to average happiness levels (e.g. Helliwell, Layard, and Sachs 2017), are frequently deemed more or less meaningless. Denmark, for example, has topped such rankings several times, but the Danes' usual response is to rush to explain that the test must measure something else, because Danes are not as happy as the report suggests (K. Christensen, Herskind, and Vaupel 2006; Booth 2015). All in all, many decision-makers desire psychometric tools, and many stakeholders reject those tools as inadequate.

### 1.2 Representing and constructing

Despite the fascinating combination of practical relevance, popularity and epistemic challenges that psychometrics embodies, philosophers of science have had relatively little to say about it. That is, until very recently. By 2018, we have a handful of philosophical analyses of the usage of psychometrics in the science of well-being (Angner 2009; Alexandrova 2017), interpretations of psychometrics in terms of the realism/anti-realism dichotomy (Hood 2008), studies of the role of modelling in psychometrics (McClimans, Browne, and Cano 2017) and more general analyses of psychometrics as an

[1] The Myers-Briggs test is in fact rejected by many trained psychometricians as invalid based on the amounting evidence. Its continued usage warrants mentioning it here as an example of the role psychometric instruments play in society.

epistemic activity (Alexandrova and Haybron 2016). In this dissertation, I contribute to the philosophical literature on psychometrics with two claims, one pertaining to representation and one pertaining to concept formation.

It might be considered a truism that measurement is representation. But what do psychometric instruments represent? What should they represent? Mathematical psychologists, and formally orientated philosophers, have argued that psychometricians fail to represent their target phenomena adequately (Krantz et al. 1971; Luce et al. 1990). This critique has been quickly dismissed, appealing to the observation that the critics are operating with the wrong conceptions of measurement and representation (e.g. Angner 2013). For example, the critics are said to have an overly demanding, formal conception of representation, which psychometricians are not interested in. In this dissertation, my first critical contribution is to argue that psychometric instruments fail *even on psychometricians' own standards of representation*. More precisely, validation of psychometric instruments fails to establish a quantitative representation of a test-independent attribute, although such a quantitative representation is typically assumed when psychometric instruments are used. Put differently, it is common to think that e.g. measures of depression represent a certain depressive state or process that exists independent of the test, although the validation of depression measures usually does not ensure the truth of such assumptions. A mismatch between what is validated and what is assumed has adverse epistemic consequences – or so I will argue.

My second contribution is more positive. It is a common complaint that concepts characterized in terms of a test operation (e.g. depression defined in terms of the Hamilton Depression Rating Scale) are inherently flawed or useless, and that consequently psychometric instruments should not measure operationally characterized concepts (Michell 1990; Green 1992; Bickhard 2001). But I will argue that it is possible to construct useful operational concepts: depression can, at least in some contexts, be defined in terms of the test that is meant to measure it. Furthermore, psychometric instruments can be legitimately employed for the study of both operationally and non-operationally characterized concepts. The important thing is to align the construction of the target concept with claims about what is being represented. If, for example, one has constructed a concept of depression in terms of the Hamilton Depression Scale, one should not claim,

or lead others to believe, that HAM-D represents a test-independent state or process. The upshot is that if we grant the legitimacy of (some forms of) operationalist psychometrics, we can appreciate and use the results of psychometric validation even when they fail to establish quantitative representations (of the kind I described above).

### 1.3 Measurement theory and concept formation

To defend these theses, I will employ two theories that have philosophical roots and that have not been previously applied to psychometrics (at least not in the way I do). The first approach is *measurement theory*: a formal theory of the conditions under which a measure adequately represents its target system. This theory goes under various names: measurement theory, mathematical measurement theory, and, perhaps most famously, the *Representational Theory of Measurement*.[2] Measurement-theoretic ideas are scattered around the literature, but the canonical statement of the approach was given in three volumes titled *Foundations of Measurement*. These books were written by psychologists Amos Tversky, David Krantz and R. Duncan Luce and philosopher Patrick Suppes, and published between 1971 and 1991.

In the few occasions that psychometrics and measurement theory have been discussed simultaneously, it has been in the spirit of juxtaposition: what are the differences between the two approaches, is one better than another, in what contexts should each be applied, are they inconsistent frameworks, and so on (especially Angner 2011; similar connotations are present in e.g. Suppes and Zinnes 1962; Krantz et al. 1971; Judd and McClelland 1998; John and Benet-Martinez 2000). I will depart from this typical comparative approach and will instead use measurement-theoretic ideas as tools for analysing psychometrics. More precisely, I will use measurement theory to study and explicate the kind of representation psychometrics aims for. To justify this usage, I will defend my reading of measurement theory against that which manifests in the more

[2] For now, I will use the term "measurement theory" rather than "the Representational Theory of Measurement", because the latter term has deep-rooted connotations that do not match my reading of the Representational Theory of Measurement. In chapter 4 I explain what reading of the Representational Theory of Measurement I subscribe to.

common, comparative discussions of measurement theory and psychometrics. The philosophical by-product is a new, fruitful explication of measurement theory.[3]

The second philosophical theory I utilize belongs to the field of concept formation, i.e. the study of how concepts are and should be defined and developed for scientific and other purposes. Conceptual engineering has its roots in the philosophy of Rudolf Carnap, but it has experienced a revival only recently (e.g. Scharp 2013; Eklund 2015; French 2015; Cappelen 2018). According to conceptual engineering, the aim of concept formation is not maximal intuitiveness (as some conceptual analysts would have it) nor is it capturing natural structures (as some natural kinds enthusiasts would have it). Rather, the aim is maximal usefulness, where usefulness can take on various context-sensitive meanings.

To my knowledge, conceptual engineering has not been previously applied to psychometrics. I will use conceptual engineering, firstly, to justify the legitimacy of operational concepts, and secondly, to explain how concepts should be formed in psychometrics. Although conceptual engineering is gaining traction, there is no unified, agreed-upon account of what it entails exactly, beyond its dedication to expedient concepts. I will therefore provide and defend my own account of conceptual engineering, thereby contributing to the growing literature on what conceptual engineering is and what it is good for.

### 1.4 Thesis structure

The structure of the dissertation is the following. Chapter 2 provides an overview of the techniques that are used to validate psychometric instruments. Chapter 3 explicates and defends my idea of the kind of representation that measurement requires. Chapter 4 argues that validation of psychometric instruments typically fails to justify the right kind of quantitative representation. Chapter 5 explicates and defends my conception of conceptual engineering. Chapter 6 applies conceptual engineering to psychometrics, arguing that we should not worry about concepts being operationally defined, but focus

[3] As I will discuss in more detail in chapter 3, my reading of measurement theory is similar to that of Heilmann (2015) and some other earlier contributors.

rather on the alignment of the constructed concepts and claims about what is being represented. Chapter 7 concludes.

## 2. Psychometrics, messy or unified?

### 2.1 What is psychometrics?

How do social scientists measure? A lion's share of any adequate answer would have to be dedicated to psychometrics.[4] Psychometric theory is wildly popular, as evidenced by the tens of thousands of publications that rely on it in the evaluation of the adequacy of social scientific instruments – not to mention all the textbooks, research centres, journals and Master's and PhD programmes dedicated to the study of psychometrics.[5] Given its popularity, it is no wonder that philosophers with an interest in social scientific measurement have taken psychometrics as a target of evaluation and critique (Suppes and Zinnes 1962; Krantz et al. 1971; Suppes et al. 1989; Luce et al. 1990; Hood 2008; Angner 2009, 2013; McClimans 2013; Alexandrova and Haybron 2016; McClimans, Browne, and Cano 2017). This chapter focuses on reviewing psychometric theory and laying the ground for appropriate philosophical evaluation of psychometrics.

There are, broadly speaking, two types of approaches to the evaluation of psychometric theory. On the one hand, there are those who treat psychometrics as a more or less unified approach, or at any rate as an approach that has a common core that most applications share, thereby making it an appropriate unit of analysis (Angner 2009; McClimans 2013; Alexandrova and Haybron 2016). Alexandrova and Haybron (2016), for example, formulate the implicit logic of psychometric measure validation in terms of theorizing the target construct, factor analysis and validation in terms of correlational evidence from other measures. They recognize that there are other types of validation processes, but treat the implicit logic as widespread across the science of well-being (Alexandrova 2017), and at any rate as representative enough to warrant treating it as a

[4] Monetary economic measures such as measures of GDP and measures of inflation are social scientific approaches that are distinct from the psychometric approach. One might also distinguish social indicators, such as the Human Development Index, from the psychometric approach, because compared to psychometric measures, social indicators undergo a different kind of validation process and are read as measures of different kinds of attributes or phenomena. Social scientists also use brain scans and other physical measures to assess social phenomena.

[5] The University of Cambridge has its own *Psychometrics Centre*, as do many other universities around the world. Many of these centres offer PhD programmes and master's programmes. Among journals that actively publish psychometric work and articles on psychometrics are, for example, *Psychometrika, Journal of Educational Measurement, Theory & Psychology, Psychological Methods, Applied Psychological Measurement* and *Assessment.* I list some famous textbooks in the next section.

unit of evaluation. Something similar is going on in Leah McClimans' paper on the psychometric validation of *Patient Reported Outcome Measures* (2013), where part of the paper contains McClimans' evaluation of what she calls the "dominant measurement methodology" in the field. She describes "the dominant methodology" in terms of the *Classical test theory*, which according to McClimans (2013, 527) "establishes validity via content and construct testing".

On the other hand, there are those more inclined to think of psychometrics as a library of techniques, theories and interpretations that get mixed and matched in applications without guidance from a unified framework. As it happens, many of the critical observers who hold this view of psychometrics have been connected to mathematical measurement theory, in particular, the Representational Theory of Measurement (RTM) (Suppes and Zinnes 1962; Krantz et al. 1971). Indeed, the messiness of psychometrics has been one of their motivators for developing mathematical measurement theories as the (allegedly) more coherent, more tenable frameworks for measurement. For example, Suppes and Zinnes (1962) explicitly motivate the need for mathematical measurement theory in terms of the unsystematic nature of psychologists' prior treatments of issues of measurement. According to Suppes and Zinnes, the psychological measurement literature is dotted with "bewildering and conflicting catechisms" rather than a unified theory:

> *While measurement is one of the gods modern psychologists pay homage to with great regularity, the subject of measurement remains as elusive as ever. A systematic treatment of the theory is not readily found in the psychological literature. For the most part a student of the subject is confronted with an array of bewildering and conflicting catechisms, catechisms which tell him whether such and such a ritual is permissible, or, at least, whether it can be condoned.* (Suppes and Zinnes 1962, 1)

While the term *psychometrics* is never mentioned, it is clearly the unsystematic nature of psychometrics that is at issue, given that psychometrics was the dominant approach to measurement at the time of their writing (and still is) (cf. Krantz et al. 1971 ch. 1, where Suppes and co-authors contrast RTM explicitly with psychometrics).

In this chapter I review classic psychometric works as well as psychometrics textbooks to give an overview of the main theories, techniques and methods that go under the header *psychometrics*. To my knowledge, this is the most extensive review of the subject in the *philosophical* literature. The review is not a summary of textbooks, but rather a conceptual presentation of themes and debates that characterize psychometric literature. In particular, I show that while the psychometric literature can be described in terms of shared, interconnected themes – notably the themes of *validity, reliability* and *model-testing* – each of these themes is characterized by a great variety of techniques and methods, and their associated rationales and interpretations. Many of the rationales and interpretations conflict with each other – witness, for example, incompatible views about the potency of *coefficient alpha* (section 2.5) and conflicting recommendations on how to respond to model misfit in *Item response theory* (section 2.4). I therefore argue that there is no Capital V Validity, no Capital R Reliability nor a Capital M Model that could be taken as the core of a unified psychometrics.[6] Hence there is no Capital P Psychometrics, but rather a host of techniques and theories, unified only by their aim – psychological testing and quantification – and intertwined roots in history.

I first provide an overview of works reviewed (section 2.2) and an overview of the three main themes (2.3). I then review this literature to show that there is no Capital M Model (section 2.4), no Capital R Reliability (section 2.5) and no Capital V Validity (section 2.6) in psychometrics. Section 2.7 concludes the review.

### 2.2 Main works reviewed

Tables 1 and 2 list the main works reviewed, where Table 1 is dedicated to what might be considered classic works in psychometric theory, while Table 2 introduces contemporary textbooks and introductions to psychometric themes. The aim of this review is to provide an understanding of key concepts, techniques and controversies in psychometrics (not a historical overview of their origins and developments), and the reviewed works have been selected to support that aim.

[6] Borsboom (2005) makes a similar claim about the diverse and non-harmonious nature of psychometric approaches to measurement. His characterization, however, is in terms of three approaches (classical test theory, latent variable models and the representational theory of measurement), whereas here I shall be concerned with differences between the first two as well as messiness that is internal to each of them.

The reader will note that in Table 1 the temporal focus is on the 1950s. This impression is further enforced when considering that I have reviewed Fiske's 1971 book, although his best-known contribution to psychometric theory occurred in 1959. Fiske's best-known paper, which he co-authored with Donald T. Campbell, is entitled "Convergent and discriminant validation by the multitrait-multimethod matrix" (D. T. Campbell and Fiske 1959). Undeniably, the works I have picked from the 1950s constitute defining works of psychometric theory. This can be seen from the fact that the selection of these works coheres with other contemporary reviews of psychometrics (e.g. Strauss and Smith 2009). The chosen works are also frequently referenced in contemporary introductions to central themes in psychometrics (e.g. Embretson and Reise 2000; Borsboom 2005; Lance, Butts, and Michels 2006).

Nonetheless, it would be unjustified to say that nothing of significance happened in psychometric theory before the 1950s. Indeed, psychometrics has its roots deeper, for example, in Karl Pearson's mathematical innovations and James Cattell's experimentations with mental tests in the late 19$^{th}$ century, both in association with Francis Galton (Mulaik 1972; Jones and Thissen 2006). *The Psychometric Society* was founded in 1935, over ten years before the publication of the earliest work I have reviewed here (Gulliksen's *Theory of Mental Tests*). The main reason for not including these earlier works is that much of their contribution is captured by the works I have included. For example, while Lee Cronbach's 1951 paper is a widely known contribution to theory of psychometric reliability, he explicitly acknowledges that his findings build on the insights of his contemporaries and predecessors. Since the aim here is not a historical overview of psychometrics, but an overview of main psychometric ideas and techniques, the review need not trace the lineage of each technique introduced. For those who are curious, I have added references to other authors' historical overviews of psychometric ideas (e.g. Mulaik 1972; Sireci 1998; Jones and Thissen 2006; Strauss and Smith 2009).

It would also be unjust to say that nothing of significance happened in psychometric theory after the 1950s, although no later decade seems to match the 1950s, neither in terms of the concentration of significant works and nor in terms of their fame. By including Messick's 1995 contribution and the multiple editions of Nunnally's (and Bernstein's) textbook, the review captures also some of the later key developments in

psychometrics. Furthermore, I have included contemporary textbooks (presented in Table 2) precisely in order to give a more up-to-date overview of psychometric theory.

| Year | Author | Work reviewed | Main themes | Reason for inclusion in the review |
|---|---|---|---|---|
| 1950 | Harold Gulliksen | *Theory of Mental Tests* | Reliability | Widely considered a classic contribution to theory of reliability. Cf. Embretson and Reise (2000, 13) |
| 1951 | Lee Cronbach | "Coefficient alpha and the internal structure of tests" | Reliability / Coefficient alpha | Coefficient alpha is the most widely used measure of reliability, and is usually known as Cronbach's alpha, due to Cronbach's classic formulation of it. |
| 1955 | Lee Cronbach and Paul Meehl | "Construct validity in psychological tests" | Explication of construct validity | Classic and defining work on construct validity, one of the first published works to explicate the concept. Cf. Strauss and Smith (2009) |
| 1957 | Jane Loevinger | "Objective tests as instruments of psychological theory" | Validity and reliability | Considered a seminal contribution to theory of psychometric validity. Cf. Strauss and Smith (2009). |
| 1968 | Frederic Lord and Melvin Novick | *Statistical Theories of Mental Tests* | Test theory | An often-cited and comprehensive exposition of developments of statistical theories of tests. |
| 1971 | Donald W. Fiske | *Measuring the concepts of Personality* | Validity and reliability | In 1959, Fiske co-authored a classic paper with Campbell: "Convergent and discriminant validation by the multitrait-multimethod matrix". Due to its narrow and technical focus, the review focuses on Fiske's more comprehensive book. |
| 1995 | Samuel Messick | "Validity of psychological assessment: Validation of inferences...." | Validity | Messick's work is widely cited as a defining contribution to psychometric theory of validity (in particular his 1989). Cf. Borsboom (2005); Strauss and Smith (2009). |
| 1967 /1979 /1994 | Jum C. Nunnally (and Ira Bernstein in 1994) | *Psychometric Theory* | Psychometric theory | Undeniable influence: widely adopted as teaching material, staggering amount of references, frequently referenced as authority on e.g. acceptable levels of reliability of a measure (Lance, Butts, and Michels 2006). |

*Table 1. Classic works included in the review.*

| **Work reviewed** | **Main themes** | **Reason for inclusion in the review** |
|---|---|---|
| Mulaik (1972): *Foundations of Factor Analysis* | Mathematical rationale for factor analysis | Helps provide an accessible overview of basic factor analytic techniques and their historical background |
| Hambleton et al. (1991): *Fundamentals of Item response Theory* | Introducing the basic concepts and tools of item response theory | Helps provide an accessible overview of item response theory and its merits compared to classical test theory |
| Kline (1998): *The New Psychometrics.* | Non-technical discussion of psychometrics and its critics | Supports the explication of various conceptions of reliability and validity |
| Embretson and Reise (2000): *Item Response Theory for Psychologists* | Foundations and practice of item response theory | Helps provide an accessible overview of item response theory and its merits compared to classical test theory |
| Rust and Golombok (2009): *Modern psychometrics: The science of psychological assessment* | Practical guide to psychometric test construction, with discussion of background and controversies relating to main psychometric concepts | Supports the explication of various conceptions of reliability and validity |
| De Vet et al. (2011): *Measurement in medicine: a practical guide* | Practical guide to choosing, developing and interpreting measures in the context of medical measurement | Supports the explication of various conceptions of reliability and validity |

*Table 2. Textbooks included in the review.*

### 2.3 Models, Reliability and Validity

What is a psychometric measure?[7] Psychometric measures typically consist of questions (or even more generally: tasks) that are posed to the test taker. For example, an intelligence test might contain questions about how to correctly continue a sequence of symbols. A test of well-being, by contrast, might ask the test taker to rate their standing on a scale from 1 (completely disagree) to 7 (completely agree) in response to the statement: "My life is close to my ideal". The questions that constitute a psychometric measure are typically called *items*. The aim of the test constructor is to devise items so that together they allow the researcher to make inferences about the test takers. For example, a good psychometric test of intelligence is such that a test taker's responses allow the researcher to infer something useful about the test taker's intelligence.

Psychometricians have a host of techniques for assessing how good the psychometric test is, that is, whether the researcher is justified in making certain kinds of inferences based on the test responses. This review has been structured in terms of three big themes that characterize such test evaluation: *models, reliability* and *validity*. The reason for picking these themes is a sociological one: psychometricians themselves consistently conceive of models, reliability and validity as central themes in the evaluation of psychometric measures. Accordingly, models, validity and reliability have a prominent place in almost all the reviewed works. Where some of these themes are missing or have a less pronounced role, the reason is the author's choice to explicitly focus on a subset of aspects of psychometrics, not a rejection of the importance of reliability, validity or models (e.g. in Cronbach (1951) the focus is reliability, in Cronbach and Meehl (1955) the focus is validity, in Mulaik (1972) the focus is factor analytic models and in Embretson and Reise (2000) the focus is on Item Response Theory or IRT models).

As the chapter will show, it is difficult to characterize the content of these themes at a general level, because each theme has been conceptualized, operationalized and linked to the other two themes in countless ways. As a first approximation, we can think of the assessment of reliability as the evaluation of properties that are internal to the measure. For example, reliability pertains to the way different questions or items on a

[7] The terms "test" and "measure" will be used interchangeably in this chapter.

single measure relate to each other. Assessment of validity, by contrast, concerns the evaluation of external properties, that is, for example, how the test of interest relates to other measures. Models, finally, are mathematical representations of relations between variables that are of interest to psychometricians, such as relations between observed test scores, measurement error and psychological attributes.

These themes have a myriad of conceptual and practical relations, some of which may *appear* contradicting. Consider reliability and validity, which many authors conceive of as closely linked. To some, reliability is a component of validity (Loevinger 1957; Messick 1995). Others argue that reliability is a necessary condition of validity, but that validity is not necessary for reliability (Kline 1998). Many point out that reliability can come at the cost of validity, that is, sometimes increasing a test's reliability decreases its validity (Cronbach and Meehl 1955; Kline 1998).[8] Still others emphasize that the important connection between reliability and validity is that together they constitute a desirable measurement property known as *generalizability* (D. W. Fiske 1971; Nunnally and Bernstein 1994). It is not important to consider (here), which of these readings of the validity-reliability link are correct or most fruitful. What is worth keeping in mind is that validity, reliability and modelling are interlinked aspects of psychometrics. Hence to structure the review in terms of models, reliability and validity is not meant to imply that there are three conceptually, let alone temporally, strictly distinct activities that together constitute psychometrics. Rather, the threefold distinction is an aid to clear explication of core psychometric activities.

Our exploration of psychometrics starts with models. This is a natural place to begin, because conceptions of reliability and validity tend to build on assumptions about structural relations between variables. In other words, notions of reliability and validity are usually grounded in implicit or explicit modelling. This, again, highlights dependencies between the three main themes, in this case models on the one hand and validity and reliability on the other.

[8] According to Cronbach and Meehl, high internal consistency may *lower* validity. Given that "certain reliability formulas describe internal consistency", as we shall see later, the former statement can be interpreted as a statement about the relation between reliability and validity.

## 2.4 No Capital M Model

### *2.4.1 Classical test theory and its critics*

The roots of Classical test theory (CTT) are in the scientific study of error and correlation, particularly in technical and conceptual innovations made in the 19th century (Traub 2005). The idea of correlation, although implicit in earlier work, was crystallized in Francis Galton's studies of heredity, while the idea of error as a random variable has its roots in astronomy. Both notions played a crucial role in psychologist Charles Spearman's 1904 paper "The Proof and Measurement of Association Between Two Things" which has been regarded as the starting point of CTT. In that paper, and some of his subsequent work, Spearman was concerned with the way random error distorts the correlation between two measures from the correlation the measures would have in the absence of error. He and his contemporaries, particularly T. L. Kelley and W. Brown, derived the main results concerning the handling of error in psychological tests, which came to form the core of CTT. CTT, in turn, was the paradigm for psychometric validation for most (if not all) of the 20$^{th}$ century and continues to be influential alongside the more recently developed psychometric approaches.

It is apparent that the concept of random error is central to CTT, because it occurs in the central proposition of the theory: the observed score is equal to the sum of two unobserved variables, random error and the *true score*. This proposition is often summarized in the form:

O = T + E
where
O is the observed score,
T is the true score on the test of interest, and
E is the error component.

This equivalence is simple, but under suitable interpretations of (and assumptions about) its components O, T and E, it is meant to allow one to derive useful indices of the influence of error on correlations between tests. These indices will be more extensively discussed in section 2.5 below. This section, by contrast, studies CTT from the

perspective of psychometric modelling. In particular, the focus is on how CTT's perceived problems have provided impetus to other frameworks for psychometric modelling.

Beyond indices of reliability, much of the methodological literature on CTT revolves around the way the psychometric test is constructed. In other words, the focus is on the rules or blueprints for constructing a total observed test score (e.g. a person's total score on an intelligence test) from scores on component questions, i.e. items (e.g. a person's score on a single question in the intelligence test). In fact, what Fiske (1971, ch. 8) calls a psychometric model is not the CTT equation described above, but rather the blueprints or rules for the construction of the test score (what Fiske calls "index"). In other words, in Fiske's terminology, a model specifies answers to the following kinds of questions: what kinds of items are there? how are items weighted to form the final total score? how is each item scored? and so on and so forth. The total score (or "index") is taken to represent "the degree of the construct […] attributed to a person" according to Fiske (1971, 137).

In Fiske's discussion, the simplest of such models is the *frequency model*,[9] where a person's score on a test containing only dichotomous items (i.e. items with only two answer categories, usually yes/no or correct/incorrect) is determined simply by counting all the times a person endorses a question – "endorse" here means simply answers yes, or correct, or whatever is taken as a positive manifestation of the target construct. No differentiation is made between items: endorsements on each item contribute equally and in the same way to the total score. The test score does not reflect any differences between two people who reach the same score endorsing completely different items.

CTT is frequently criticized for a lack of emphasis on the characteristics of the items. That is to say, CTT is criticized for not taking into account that some items are e.g. more difficult than others, and that this has a significant bearing on the interpretation of test results (e.g. McClimans 2013). However, in what Fiske calls the *cumulative homogeneity model* (which Fiske takes as an improvement on the frequency model, broadly speaking) the items are ordered with respect to the way they manifest the target

[9] Fiske's usage of the term "model" may seem unorthodox, because he is discussing the design of the structure of the test, not relations between variables. I shall respect Fiske's usage of the word in this section.

attribute. For example, in an arithmetic test the items are ordered according to their difficulty.[10] In the frequency model, the difficulty of the item is established by examining the frequency with which subjects endorse each item, under the assumption that harder questions are endorsed less frequently than easy questions. The total score of the subject is again simply the sum of endorsements. But the score, and the instrument underlying it, is only taken to be adequate if the subject's ordering in terms of total scores "matches" the ordering of the items in terms of difficulty. In other words, a score (and the instrument) is taken to be a good indicator of the target construct, if people with low scores tend to get easy questions right and fail most of the difficult ones, while people with high scores tend to answer both the easy and the difficult questions right. Thus, on this approach, an ideal instrument is such that when the items are ordered from low to high in terms of difficulty, each subject has a string of endorsements followed by a string of items that are not endorsed.

Both the frequency model and the cumulative homogeneity model (if we follow Fiske in calling them models) can be thought of as grounded in CTT in the sense that in both approaches the subject's observed score is taken as a relatively direct indication of the subject's true score. Characteristics of the items, e.g. difficulty, are not used in the estimation of subjects' standing on the target attribute. But it is useful to note already that the researcher who uses the cumulative homogeneity model nonetheless *considers* features of the items. The researcher using the cumulative homogeneity model ranks items in accordance with their difficulty and uses this information to determine how good the test and its individual items are. To assess this aspect, Fiske develops correlational techniques for determining how well the model is fulfilled.

With this background in mind, let's move on to criticisms of CTT. One of the most prominent criticisms of CTT concerns the way the difficulty of the items is determined in CTT (Hambleton, Swaminathan, and Rogers 1991; Kline 1998; De Champlain 2010). The problem with CTT, according to critics, is that since the difficulty of the test items is

[10] Here is another interesting example from Fiske (1971, ch. 8): in a test of *willingness to disclose personal information,* the items could be ordered in terms how "difficult" it would be to endorse the item, i.e. how open you would have to be to be likely to endorse an item. For example, the item "I would tell a casual acquaintance about my most shameful thoughts" is harder to endorse than "I would tell a casual acquaintance my dog's name".

evaluated in terms of the group taking the test, i.e. as the proportion of examinees in a group of interest that answer the item correctly, the estimates of the difficulties of the items are "group-dependent" (De Champlain 2010). That means that the test will appear difficult if the group consists of low ability subjects and easy if the group consists of high ability subjects. Similarly, ability estimates derived from test scores are always "test-dependent". That means that if a test is difficult, the subjects will appear to have low ability, and if a test is easy, the subjects will appear to have high ability. Thus, the difficulty of the test varies when the abilities of the subjects in a group vary, and estimates of the abilities of the subjects vary when the difficulty of the test varies.

The upshot of these two dependencies is that it is very difficult to compare i) ability estimates (of the same ability) obtained from different tests, and ii) difficulty estimates obtained from different subject groups. Suppose one has two tests, A and B that are meant to examine the same ability, and two groups that have each taken a different test. How could one compare the abilities of two individuals from different groups, when their respective scores are on different scales (because they come from different tests) and the functional relationship between those two scales is unknown? Say Anne takes test A and obtains a score of 32, and Bob takes test B and obtains a score of 16. If test B is much harder than test A, Bob might have an equal or higher ability than Anne, even though Anne's observed score is much higher than Bob's. To obtain a meaningful comparison of the abilities of Anne and Bob, then, one needs to know the relation between the two scales. To know that, one needs to know the relative difficulties of the two tests, that is, one must be able to compare the two tests in terms of their difficulties. But the difficulties of the tests are also not easily comparable, because difficulty is group-dependent, and different groups have taken different tests. To compare the difficulties, one needs to be able to compare the abilities of subjects from different groups. But that is just the problem we started with – how to compare the abilities of two individuals from different test groups! The circularity is vicious.

Test-/group-dependency is one of the grounds on which CTT has been critiqued. In addition, a very common criticism of CTT is that its central equivalence, O = T + E, is either unfalsifiable or obviously false, depending on one's interpretation of the components of the expression. What interpretations are there for E and T? E is usually

defined as random error that has a mean of zero and is uncorrelated with true scores (Gulliksen 1950, 6-7; Lord and Novick 1968, 56). Random error is most easily explained via its contrast to systematic error. For example, error is systematic if the test consistently under-scores candidates, that is, test takers receive scores that are consistently below their true score. Error is random if deviations are equally likely to occur below and above the true score. Furthermore, an error is categorized as systematic if the observed scores for people with a high true score always contain less error than the observed scores for people with low true scores. The error is random if it is not thus patterned in accordance with true scores.

Definitions of T, the true score component, are more controversial. There are at least two widely used but obviously non-equivalent definitions of true score – Lord and Novick (1968, sec. 2.1) introduce three and Nunnally and Bernstein (1994, 211 and 217) argue that there are many more. On the one hand, true score has been taken to refer to *true ability*, that is, the degree of the target attribute the test subject possesses (e.g. Gulliksen 1950, 4). In other applications, on the other hand, true score has been defined as "the average of all scores a person would receive if he or she took the test an infinite number of times", or less demandingly, if he or she took the test repeatedly over a large number of occasions (Streiner 2003, 99). In sum, there are two prominent alternatives, which I shall call the "true ability" interpretation of true score and the "repeated administrations" notion of true score.

Depending on the true score interpretation one endorses, different problems come to view. On the true ability interpretation, the main problem is that (when paired with the assumption of random error) the equation O = T +E is implausible: the observed score is unlikely to be the simple additive function of a subject's standing on the underlying attribute of interest and random error. Rather, the observed score is likely to deviate from the true attribute levels in systematic ways. For example, if an ability test is written in a language that is not the native language of some of the test takers, this will tend to systematically distort the results toward lower scores independent of test takers' true ability (Streiner 2003). Similarly, if (some) test takers are cheating, or tired, or disinterested, or trying to give answers that please the researcher, the observed scores will reflect

something in addition to true ability and random error, i.e. they will reflect tiredness, disinterestedness, proneness to cheating and so on.

On the repeated administrations approach, one pressing question concerns the reasons for why one should care about the expected score on infinite administrations of the test. Isn't it more important to know about the relationship between observed score and the attribute of interest rather than the relation between the observed score and the average of observed scores on repeated tests? Another problem is that it is often misleading to label the divergence between observed score and true score as *random* error, because in repeated administrations of psychological tests people's responses tend to change in systematic ways, such as due to learning, memorizing, tiring and so on. The idea that observed score is the function of the average of repeat administrations of a test and random error is therefore implausible.

In some texts, such as Lord and Novick 1968, this last objection has been dodged by *defining* the relevant repeated test administrations as independent of each other, such that scores are not determined by learning, memorizing and other effects. In other words, rather than speaking of actual repeats of a test, in which attempts are likely to depend on each other, repeated administrations are defined in terms of *thought experimental test-administrations* that are independent by construction (for a detailed analysis, see Borsboom 2005). By these means, the problem that O = T+E is unlikely to be true is diverted by rendering it true by definition, that is, unfalsifiable. But this strategy looks suspicious, because who cares about an unfalsifiable theory?

The various problems I have noted have been summarized in Figure 1. These problems are often cited as reasons for moving on to another framework for psychometric testing, the Item response theory or IRT framework. In the next section, I will introduce IRT and briefly look at how it is meant to improve upon CTT.

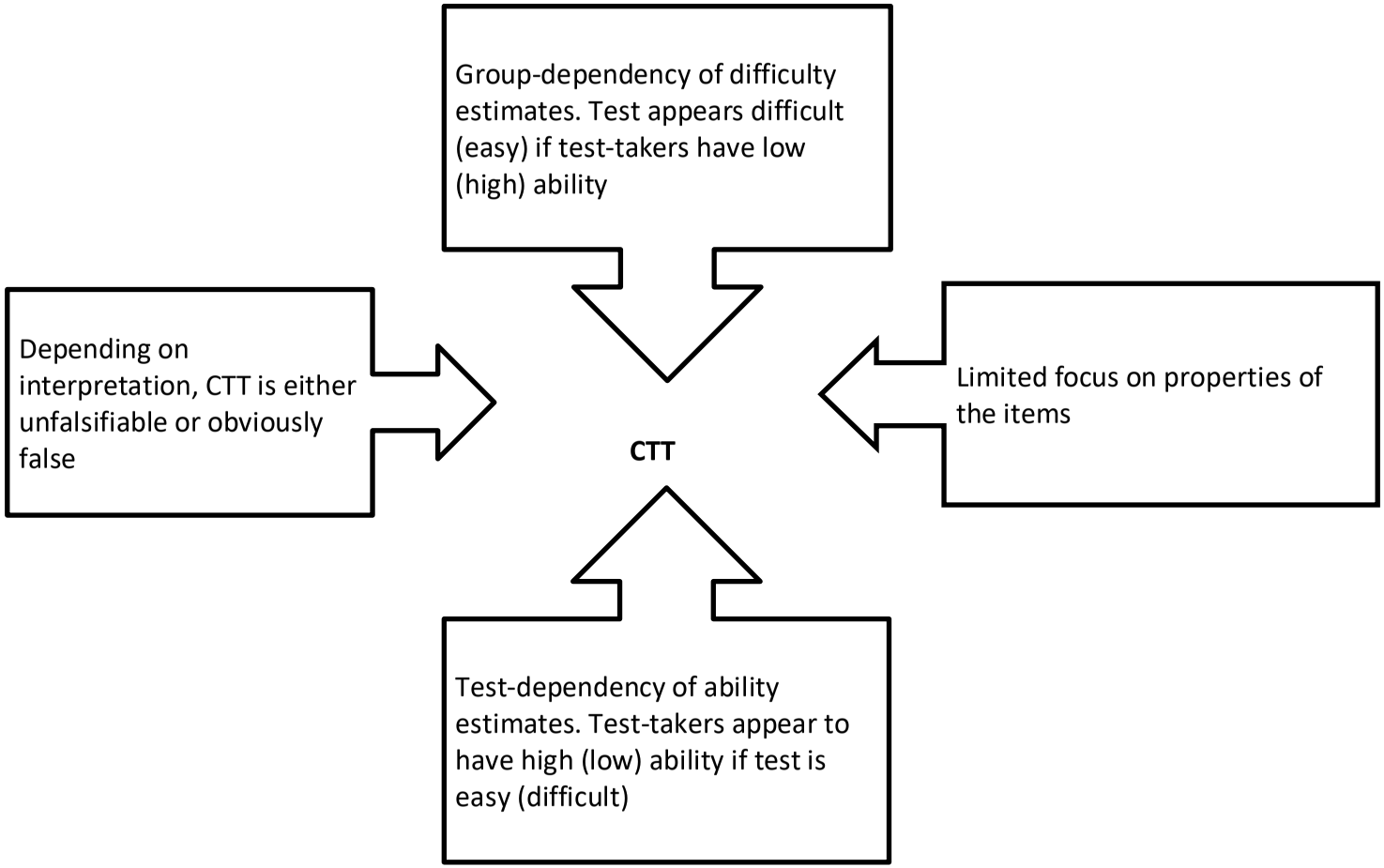


*Figure 1. Perceived problems with CTT.*

### *2.4.2 Item response theory*

IRT, or modern test theory (Nunnally and Bernstein 1994), or model-based measurement[11] (Embretson and Reise 2000), or latent trait theory, corrects for many of the shortcomings of CTT. Broadly speaking, IRT models explain the probability of a correct response (or the probability of a specific response, if there are multiple response categories per item) as a result of the influence of the ability level (or more generally: standing on the attribute) of the examinee and characteristics of the test items, where "characteristic" denotes properties like the difficulty of the test item.

The Rasch model is the simplest of IRT models: it includes one item characteristic, namely, difficulty, while other models include parameters for characteristics such as item discrimination (how informative the item is of different ability levels) and susceptibility to guessing (how likely low ability examinees are to give a correct response due to guessing). The more complex models are often needed, because often test performance cannot be explained (or predicted) only in terms of the ability levels of the test takers and the difficulty of the items. If the response data does not "fit" the simplest model, other, more complex models may be examined. Such an approach highlights a crucial difference between IRT and CTT: IRT models are clearly treated as falsifiable (Hambleton, Swaminathan, and Rogers 1991, 7).

There are two widely used expressions of the Rasch model, one that treats probability of the correct response as a dependent variable and one that treats log-odds of a correct response as the dependent variable. The first expression of the Rasch model specifies the following relation:

$$P_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}$$

where

$P_i(\theta)$ is the probability of a correct response to item i from a randomly selected examinee whose ability level is $\theta$, and

[11] The notion of model-based measurement is important in recent philosophical discussions of measurement (e.g. Tal 2016). These will be discussed in later chapters and are not implied here.

$\beta_i$ is the item difficulty parameter.

When the model is used in psychometric practice, the data that is collected using the psychometric instrument of interest is tested against the Rasch model. That is, various tests are imposed on the data to check whether it exhibits the properties that the Rasch model postulates.

One of the first steps in validating a test in terms of the Rasch model is estimating item and ability parameters from the data. The goal is to estimate item difficulty and subjects' abilities independent of each other, to avoid the test-/group-dependency problem that plagues (some applications of) CTT. But the difficulty for the IRT modeler is the same as that encountered by those working within CTT. To infer the test-independent ability of the subject from their manifest score, we need at least some estimate of how difficult the items are. If we just look at how many correct answers a subject gives relative to other people in the test group, we end up with ability estimates that do not reflect the difficulty of the test and that therefore do not generalize beyond the test. Similarly, to infer something about the group-independent difficulty of the item from the way subjects responded to that item requires that we know something about the abilities of the subjects.

The solution that has been proposed in the literature, and that practicing IRT modelers utilize, is reliance on the epistemic benefits of iteration.[12] One popular method of implementing iteration is the *joint maximum likelihood method* – thus called because item and ability parameters are estimated jointly. There are many other methods, but I will briefly describe this popular one. Roughly speaking, the joint maximum likelihood method estimates the ability of an examinee by investigating the likelihood of her response pattern (correct/incorrect on each item) conditional upon different levels of ability. The same kind of estimation is done for the item difficulty parameter: the difficulty of the items is estimated by investigating the likelihood of the manifested response patterns conditional upon different levels of difficulty.[13] The process is iterative, in the sense that the estimation

[12] The epistemic benefits of iteration are widely celebrated in recent historico-philosophical measurement scholarship. See Chang (2004, 2017a).

[13] Here is a more detailed description of the ability estimation step – the item estimation step "mirrors" this and can be understood vis-à-vis ability estimation (Embretson and Reise 2000, 201). Assume we have some

of the two parameters is done multiple times, building on information from previous estimation rounds. Thus, one might start from crude estimates of difficulty (based, say, on the frequency of correct responses), use those difficulty estimates as the tentative reference in the likelihood estimation for ability levels, use those ability estimates as the background for re-estimating difficulty levels, and so on and so forth. The ability and difficulty estimates keep updating in these consecutive steps, and the iterative re-estimation is carried out until neither estimate changes in two consecutive iteration steps.

Once the parameters have been estimated, one proceeds to test the fit between the Rasch model and the data. To that end, a battery of statistical tests is imposed on the estimates that the joint maximum likelihood method yields. For example, one goodness-of-fit test divides the examinees into ability groups based on the ability estimates and compares the observed response patterns of each group to the predictions of the Rasch model (see Hambleton, Swaminathan, and Rogers 1991, 59; Embretson and Reise 2000, ch. 9). In simplistic terms: one plugs in a given ability value *a* and a given item difficulty level *d* into the Rasch model and then proceeds to check whether the thus computed probability of correct response matches the actual frequency with which individuals of ability *a* gave the correct answer to an item that is of difficulty *d*. If the data fits the model reasonably well in the sense that the predictions of the model and the observed response patterns converge, that is taken as evidence that the attribute of interest has the attribute structure postulated in the Rasch model. Importantly, lack of fit between the model and the data means that either the test is not appropriate or the hypothesized model of the target attribute (here, the Rasch model) is not appropriate. In the latter case, other models might be tested against the data. Here, again we can note a

---

rough estimates for difficulties of items. Given an examinee who responded "correct" on a specific item, we can find the likelihood of a correct response to that item, conditional upon different ability values. (We can read these likelihoods off the so-called item response curves, which express the likelihood of a correct response as a function of ability levels. These have been preliminarily fixed on the basis of some crude characterization, say, the frequency of that item being endorsed. See chapter 4 for details.) We do this likelihood estimation for all levels of ability. This procedure is repeated for all the items. Then we can estimate the examinee's ability level by choosing the value of the ability parameter that maximizes the likelihood value of the response pattern for that examinee. In procedural terms, we choose an ability level, find the likelihoods for the specific responses our examinee gave to each item and multiply those likelihoods. Then we repeat this procedure for all possible ability levels. From the resulting likelihood values for a response pattern, we pick the ability level that maximizes the likelihood of the observed response pattern.

difference to CTT: when a model does not fit (after a sufficient number of attempts), it is considered falsified (for that context) and other models are tested instead.

| **CTT** | **IRT** |
|---|---|
| Group-dependency of difficulty estimates. Test appears difficult (easy) if test-takers have low (high) ability | Item difficulty conceptually group-independent* |
| Test-dependency of ability estimates. Test-takers appear to have high (low) ability if test is easy (difficult) | Ability conceptually test-independent* |
| Depending on interpretation, CTT is either unfalsifiable or obviously false | Proposed models of test behaviour are falsifiable (and frequently rejected if false in context C) |
| Limited focus on properties of the items (although difficulty is frequently taken into consideration) | Item properties, such as difficulty, discrimination and susceptibility to guessing taken into consideration |

*Table 3. A comparison of features of CTT and IRT, with an emphasis on the shortcomings of CTT that IRT is commonly thought to correct for.*

**In practice, estimates of ability (difficulty) inform the estimation of item difficulty (ability) in the joint maximum likelihood method. The crucial difference to CTT is therefore the conceptual innovation of not having to define ability (difficulty) in terms of the properties of the test (properties of test takers).*

### *2.4.3 IRT versus CTT – a difference in aims?*

We have seen that IRT responds to many problems that critics have pointed out with the CTT approach. These problems, and their proposed resolutions in the IRT framework, are summarized in Table 3. Why is CTT still used, if IRT is known to be superior?

Let's first note that CTT psychometricians have not just sat on their hands or looked the other way in the face of these critiques, but a multitude of techniques for dealing with the challenges has been proposed within the CTT framework. We already saw Fiske's attempts to construct indices that would reveal whether subjects' ordering in terms of total scores "matches" the ordering of the items in terms of difficulty, thereby incorporating item characteristics into the CTT framework. Gulliksen (1950, 367-371) and Lord and Novick (1968) characterize multiple techniques for analysing and overcoming

(what I have called) group-dependency of the test. Embretson and Reise (2000, 14-15 and 21) also mention techniques within CTT that deal with the test-dependency of ability estimates as well as methods that relax the traditional, strict assumptions concerning error. Given the simplicity of the CTT framework, and the ease of understanding and employing it, it is not surprising that its users have not abandoned it for IRT, but rather sought to resolve issues within its bounds.

Beyond these improvements, a fair assessment of the alleged superiority of IRT over CTT would have to consider their respective aims. Under a particular interpretation of CTT, which I will call "test focused CTT", IRT and CTT turn out to have different aims. The difference in aims, in turn, undermines some of the criticisms levelled against CTT.

Unlike IRT, test focused CTT does not seek what underlies data, i.e. it does not seek the correct model of how an attribute and item characteristics interact to determine the observed score. The equation O = T + E is not a hypothesis about what underlies a test score, but a definition – this is in fact what Gulliksen (1950, 5) says explicitly and what is continually emphasized in Lord and Novick's ground up construction of the classical model (1968, ch. 2). Consequently, true score is not defined as the underlying, test-independent attribute that the observed score manifests, but rather true score is defined as the expected score on infinite, hypothetical administrations of the test of interest. When true score is defined in terms of the expected score on infinite hypothetical administrations of test t, and error is defined in terms of the random perpetuations that lead the observed score to diverge from the expected score, the CTT model is true by definitions of its components.

So far this defence of CTT may sound like trivial hocus-pocus with the sole aim of rendering the proponent of CTT correct come what may, by his own definitions. But taking this approach to CTT, the emphasis is on reflective construction of the test rather than testing whether the model is correct – the latter aim would indeed be silly. Test focused CTT usage aims to construct a test such that the true score *qua* expected result on infinite hypothetical administrations of that test is what the researcher is ultimately interested in – it aims to make the test such that true score *qua* expected score on that test is worth investigation. The focus is on an interesting test, not on an interesting model.

With this focus, knowledge of the properties of the test as a whole and its individual items becomes crucial. The various reliability coefficients that have been derived from the complex network of definitions and assumptions CTT is embedded in (which we will explore more in section 2.5) are used as information about how well the observed score gets at the true score, that is, how much of that undesirable error a certain test construction leads one to have to endure. Analyses of the properties of the items, such as the kind of tests of difficulty and group-dependence mentioned in the beginning of this section, can be used to understand the properties of the test as a whole – for example, why the test as a whole appears to have low reliability. This knowledge then feeds into better test construction. In the words of Lord and Novick (1968, 327):

> *[T]he statistical characteristics of the total test depend entirely on the statistical characteristics of the items used to build it. […] Knowledge of the item characteristics and their effects helps us understand and allow for the peculiar measurement properties of a particular psychological test. Item analysis may enable us to construct tests with specified or, in a limited sense, optimal measurement properties.*

If we adopt this reading of CTT, the criticisms that CTT is unfalsifiable or obviously false must stem from a misunderstanding of what CTT aims for. Plausibly, the criticisms are rooted in an endorsement of IRT's aims and an extrapolation of those aims to the evaluation of CTT. Of course, there may be instances of CTT usage where the researcher has aims that are similar to the goals that characterize the IRT approach. This is *prima facie* plausible, because the theoretic background of CTT is rather complex and its terminology (e.g. true score) invites misinterpretations. In any case, I have tried to show that IRT is not necessarily a straightforward improvement on CTT, but rather a whole different outlook on the role of models in psychometric test construction.

For the sake of completeness, it must be added that different aims exist also within IRT – it is not all model-testing with the aim of finding the best-fitting model. Several theorists that are committed to the primacy of the Rasch model over the other, more complex IRT models, argue that if the Rasch model does not fit the data, it is not the *model*

that should go, but rather the *data* and the *test* that yielded the data. The idea is that since the Rasch model has such desirable properties – properties that some argue are necessary for measurement and that other IRT models do not manifest – psychometricians' aim should be to devise tests that yield data that fit the model. This approach evidently contrasts with the previously described model-fitting approach: one aims to find the model that fits the data, the other aims to find the test that yields data that fits the model (Andrich 2004).

Pulling all these observations together, we have seen that IRT has a model-fitting and a test-fitting arm, and that CTT comes with a multitude of interpretations. And we have not even began to touch upon the latest and most advanced IRT based models that have such futuristic names as the *IRT-ZIP model*. Nor have we looked at the various coefficients of reliability that have their grounding in CTT, or the modifications of the basic model (Lord and Novick 1968). The upshot is that there is much beyond and in between the two allegedly opposing psychometric frameworks.

### *2.4.4 Factor analysis*

A discussion of psychometric models would not be complete without mentioning *factor analysis*, a frequently used technique. As with IRT, factor analysis looks for models of latent variables that best explain the observed test behaviour. Unlike IRT, though, factor analysis takes *correlations between test items* as the explanandum, rather than analysing patterns of responses on individual items. In factor analysis, the latent (as opposed to observed) variables are discussed in terms of *factors* – the concept of latent ability and factor do come apart in important respects, but for now we can keep treating them without making those differences explicit.

While factor analysis has its roots in Francis Galton's studies on inheritance in the late 19$^{th}$ century, and Karl Pearson's involvement in those studies, many of the technical developments occurred in relation to the study of intelligence in the early 20$^{th}$ century psychology (Mulaik 1972). Prominent psychologists such as Charles Spearman (credited for the first factor analytic model), Godfrey H. Thomson, Cyril Burt and L. L. Thurstone were interested in the nature of the ability (or intelligence) that manifests in test behaviour. In particular, their questions pertained to the generality and differentiation

of the ability or abilities involved in various tests, for example tests pertaining to mathematics, English, classics, music and so on. Is there one very general ability that drives all test behaviour? Are there a few different kinds of abilities, each of which is involved in some subgroup of tests? Or is mental ability even more differentiated, so that in each test a sample of abilities is involved (without any clear matching of groups of tests with some specific ability)? In other words, if the model of the observed test result of some test t is written as

$$O_t = w_{t1}A_{t1} + w_{t2}A_{t2} + w_{t3}A_{t3} + \cdots + w_{tn}A_{tn}$$

where $O_t$ is the observed variable on test t,

$A_{t1}…A_{tn}$ are latent abilities, and

$W_{t1}…w_{tn}$ are weights,

then the question these psychologists asked was which abilities (if any) were such that they occurred i) in models of all tests, ii) only in models of some subgroup of tests, and iii) only in models of specific tests. The various techniques of scrutinizing test correlations that later became known as factor analytic were developed to uncover these underlying factor structures.

In the literature on factor analysis, a distinction is made between *exploratory* and *confirmatory* factor analysis. As the name suggests, in confirmatory factor analysis the researcher first proposes a hypothesis of the kind of factor structure she expects, on some theoretical grounds, to be involved in the tests she is interested in. She then uses factor analysis to determine whether some observed data fits her expectation, that is, whether her hypothetical factor structure explains the actual data. In exploratory factor analysis, by contrast, the researcher has no prior theoretical commitments, but rather various factor structures (e.g. different numbers of factors and different weights) are explored to find the best fitting model. The choice of number of factors and the possibility to try different weights for these factors allows one to generate and explore many models.

To provide further intuition of how factor analysis works, let us consider a simplified version of the factor analysis process. Consider a test situation where subjects

see various descriptive adjectives such as “energetic”, “happy”, “gloomy” and “anxious” and are asked to rate their standing in terms of these adjectives using a scale from 0 to 3, where 0 stands for “not at all” and 3 stands for “very much”.[14] Given responses, the researcher then proceeds to calculate correlations (Pearson correlation coefficient) between each pair of items (in this case each adjective is an item). She then either lets her theory determine the number of factors in terms of which she analyses the data, or she leaves the number open for exploration. Let us, for the sake of illustration, concentrate on a two-factor case. From the matrix of correlations, the researcher calculates (that is, runs a statistical test to determine) what the weights for each of the two factors should be in the model for each specific item, given how that item correlates with all other items. For example – and here I am speaking somewhat vaguely, because the intuitions concerning pairwise comparisons might not exactly match what happens when the whole matrix is analysed – if items “happy” and “pleased” are strongly positively correlated, then they are not just likely to share a common factor (or factors) but weights on that (or those) common factor(s) are likely “similar” (e.g. in this case both have a similar loading especially on the factor represented on the x-axis.) The resulting weights are known as factor loadings.

In the left graph of Figure 2, the axes represent the two factors, each point represents an item (adjective in this case), and the position of each point reflects the way that item loads on each factor. For example, the item “strong” has a loading of roughly 0.5 on the factor represented by the x-axis and a loading of around 0.25 on the factor represented by the y-axis. It is clear that the items form two clusters, but the clusters do not correspond to the two factors, that is to say, both factors are needed to explain or account for the clusters. However, if we tilt the axes slightly, they pass closer to the clusters. The red lines on the left represents such rotation, and the resulting situation is depicted on the right. The two clusters now have a stronger association with their respective factors, so that items in one cluster have a high loading on “Factor X” and a low loading on “Factor Y”, and vice versa for the other cluster. These loadings have been

[14] This example is based on items that I have hand-picked from the 72-item Motivational States Questionnaire. The data set used is available in the psych-package in R Studio. The correctness of the data set is not of interest in this chapter, because the point is not to generate genuine, interesting factor analytic results but rather to illustrate how factor analysis works. For more on the data set, see https://personality-project.org/r/html/msq.html

achieved with the *varimax-rotation*, but there are many others, each of which has their associated rationales.

The practice of axis rotation is a core activity in factor analysis, but for the uninitiated it might strike as odd. In many cases such a rotation would amount to a distortion of the data – for example, if we plot children's weights and heights on such a graph based on a good data set and then rotate the axes in a similar fashion as above, the result is clearly a wrong depiction of the relationship between children's weights and heights. However, in many of the cases factor analysts are interested in, the axes do not have a straightforward interpretation in terms of well-delineated attributes such as weight and height. Rather, the axes are simply aiding in visualizing the loadings that arise when test scores undergo the above described statistical testing. An interpretation of the axes in terms of attributes is external to the statistical analysis, although the rotation of the axes is often thought of as a tool for investigating the tenability of various interpretations. For example, if we have some independent theoretical reason to expect that *one* latent attribute explains items "gloomy", "anxious" and so on and some *other* latent attribute explains items "energetic", "happy" and so on, then in light of the rotation depicted on the right, the data can be thought of as providing support for that hypothesis.

Of course, while the openness of the meaning of the axes allows for the handy trick of rotation, it also means that from a purely statistical perspective, all the rotations are equally permissible or valid. Hence there is much room for debate about whether a given factor analytic result supports a theory or not, and a wide range of incompatible hypotheses can be supported by the same data just by switching the rotation. This explains why it can be confusing to equate factors with latent abilities: latent abilities often connote rich meanings and concrete denotations (something akin to the well-delineated attributes of weight and height) while a factor may simply be a statistical object without a concrete denotation. It is perhaps most appropriate to say that the (typical) aim of factor analysis is to discover factors that correspond to latent attributes, but one should be careful not to treat every factor structure that fits data as corresponding to a latent attribute structure, without further arguments.

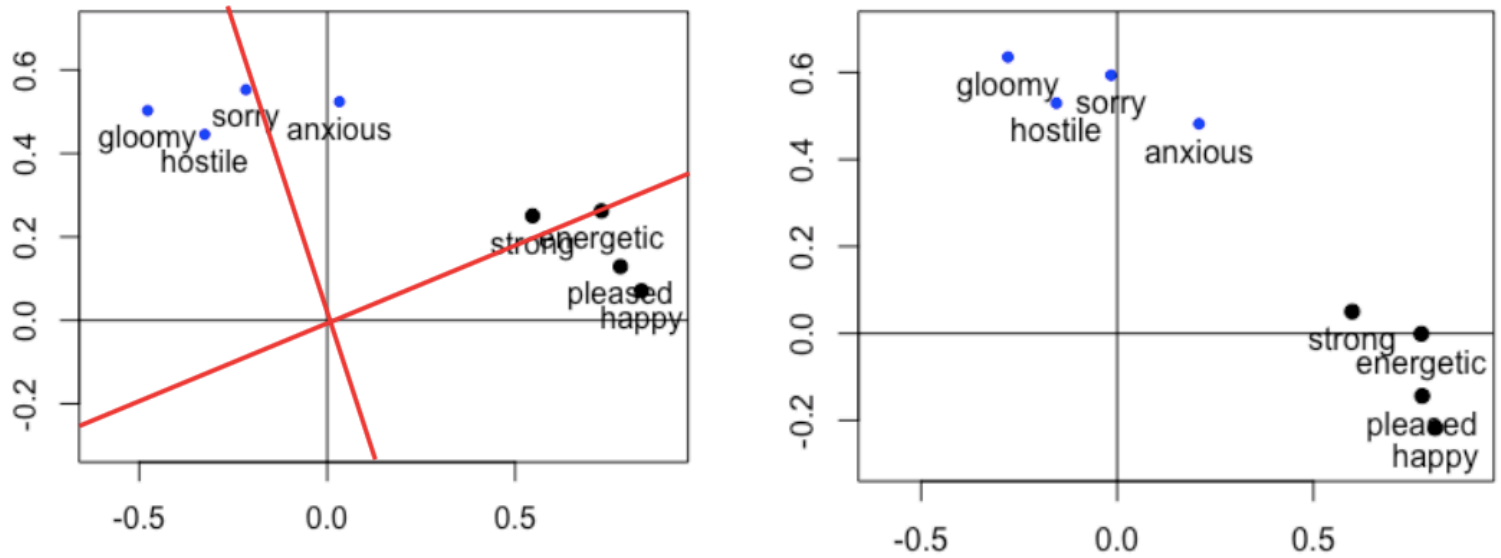


*Figure 2. Example of rotation in factor analysis. The plot on the left is unrotated. The plot on the right is the result of varimax rotation. The red lines indicate how the plot on the left rotates to the plot on the right.*

### *2.4.5 So many models*

Table 4 recaps the variety of modelling exercises that characterize psychometrics. Evidently, there is no single Capital M psychometric model.

| | **Classical Test Theory CTT** | **Item response Theory IRT** | **Factor analysis** |
|---|---|---|---|
| **What kinds of models?** | O = T + E<br>Observed score O as a function of true score T and error E* | $P_i(\theta) = f(\theta, \beta_1 \ldots \beta_n)$<br>Probability of correct response $P_i(\theta)$ as a function of ability $\theta$ and item parameters $\beta$ | $O = w_1F_1 + \cdots w_nF_n$<br>Observed score O as a function of factors $F_1 \ldots F_n$ and weights $w_1 \ldots w_n$† |
| **How are models used?** | Mainly as a definition or assumption on which indices of measure properties are built | Attempts to find the model that fits the data, or the test that fits the model (some branches of Rasch analysis) | Attempts to find the model that best accounts for correlations between tests |
| **How many model variants?** | There are many interpretations of the core component T | Many different variants, e.g. different number of item characteristics | Basically, infinitely many models may be explored by varying number of factors, weights and form of their relations |
| **Is it still used?** | Yes | Yes | Yes |

*Table 4. Summary of some of the main modelling approaches in psychometrics.*

**Much focus has been laid on models qua rules for the construction of total score from item performance.*

*†There is also non-linear factor analysis*

## 2.5 No Capital R Reliability

Classical Test Theory is the foundation of the psychometric approach to reliability. Beyond that, conceptions of reliability vary. Reliability has been variously described using terms such as accuracy, repeatability, consistency, representativeness, (lack of) variability, (lack of) error, stability and homogeneity. Consider the examples in Figure 3.

All the quotes depicted in Figure 3 are from contexts in which the author is engaged in an account of reliability in psychometric theory. Moreover, as is apparent from many of the quotes, the authors present themselves as saying something that is "widely accepted" within psychometrics. Yet, the characterizations differ. Are these authors characterizing the same thing with different words, or characterizing different things with the same term ("reliability")?

While the quotes testify that the conceptualizations of reliability differ across the reviewed texts (at least on the surface), it is common for reliability theorists to describe two classes of tests for the assessment of reliability. These are the *test-retest reliability* and the *parallel forms reliability*. I shall go over these ideas one at a time, aiming to uncover whether there is a shared Capital R Reliability.


*Figure 3. Quotes about reliability. Sources from left to right, starting from top left: Cronbach 1951, 297; Lord and Novick 1968, 139; Kline 1998, 26; Rust and Golombok 2009, 72; Fiske 1971, 153; Loevinger 1957, 636; Nunnally and Bernstein 1994, 248; Cronbach, Rajaratnam, and Gleser 1963, 144. Emphases added.*

### *2.5.1 Test-retest reliability*

Test-retest reliability is exactly what it sounds like: a test of whether or not, and to what extent results on two administrations of the same test converge. Loevinger

(1957), Fiske (1971) and Nunnally and Bernstein (1994) all call this kind of reliability "stability". Kline writes that:

> *[Test-retest reliability] is an essential attribute for any good measure, whether psychometric or not. A test should yield the same score for each subject when he or she takes the test on another occasion, given that their status on the variable has not changed.*

There is, broadly speaking, agreement among the authors on the meaning and testing of test-retest reliability. We get a hint of this, for example, from the fact that many of the reviewed texts merely mention stability but dedicate much more space for the explication of other types of reliabilities – evidently test-retest reliability does not need such extensive elaboration (Fiske 1971, 152; Nunnally and Bernstein 1994, 212; Kline 1998, 29). There also appears to be agreement that high test-retest reliability is not always desirable in practice. For example, according to Nunnally and Bernstein stability (in the present sense) "may or may not be" desirable, depending on what is being measured (1994, 212). By this they mean, presumably, that while stability of measurement results is crucial if the target phenomenon is stable, such stability is so incredibly rare in psychology that the requirement of stability of tests results is impracticable (Cronbach 1947 articulates this point concisely). Loevinger (1957) also argues that what one should conclude from stability or lack thereof depends on the situation. There can be "spurious stability" as in when the subject's standing on the target attribute has changed but her observed score remains the same, and there can be "spurious change" as in when the subject's standing on the attribute is stable, but the scores have changed. In light of such spuriousness, the requirement of high test-retest reliability is not appropriate across measurement contexts. Thus, while the characterization of test-retest reliability is shared, we would be hard-pressed to insist that test-retest reliability is the Capital R Reliability unifying all psychometric testing.

### *2.5.2 Parallel forms reliability*

There is much more to be said about the parallel test reliability. At the informal level, there is again a broad-based agreement on what is at issue: parallel test reliability (which goes under several names) concerns an assessment of the extent to which two comparable tests correlate (Gulliksen 1950, 13; Loevinger 1957, 676-677; Cronbach, Rajaratnam, and Gleser 1963; Fiske 1971, 152; Nunnally and Bernstein 1994, 212; Kline 1998, 30). Speaking broadly and informally, Fiske (1971, 152) summarizes the relevant concern thus: "How much difference did it make that we used this particular set of [test] items rather than some other set obtained by the same procedures?"[15] This much might be agreed upon, perhaps, but the challenges to a unified approach begin as soon as we look deeper into the theory underwriting parallel test reliability, as well as the definitions and tests that are associated with that theory. I shall introduce the conflicts in two steps. I first give a theoretical explanation of reliability (relying largely on Gulliksen (1950) and Lord and Novick (1968))[16] and introduce the most common method of assessing reliability (due, arguably, to Lee Cronbach). I will then explain some of the various conflicts and debates vis-à-vis these descriptive accounts.

In the theoretical literature, what I previously called comparable tests (following Fiske 1971) are known as *equivalent tests* or *parallel tests* (Gulliksen 1950; Cronbach, Rajaratnam, and Gleser 1963). We find, for example, the following definitions of reliability:

> *For the present we shall define reliability as the correlation between two parallel forms of a test.* (Gulliksen 1950, 13)

> *The classical approach defines reliability as the correlation between equivalent measures.* (Cronbach, Rajaratnam, and Gleser 1963)

[15] There are no universal, fixed procedures for the formulation of items. But there are guidelines (e.g. de Vet et al. 2011, sec. 3.4).

[16] These are considered milestone books on theory of reliability. See e.g. Borsboom (2005, ch. 2) and Walsh (1968).

Two tests are parallel if "it makes no difference which test you use" (Gulliksen 1950, 11). To get at the more rigorous definition, we need to recall the central equivalence of CTT:

O = T + E

where
O is the observed score,
T is the true score on the test of interest, and
E is the error component.

In light of this background, we are able to grasp Gulliksen's more rigorous, twofold definition of parallel tests:

> *The true score of any person on one test must equal the true score of that person on the other parallel test* (Gulliksen 1950, 11), and
>
> *The standard deviation of the errors on one test is equal to the standard deviation of errors on the other test.* (ibid., 12)[17]

This twofold definition of parallel tests[18] has been chosen so that observations of results on parallel tests open the door for claims about what the developer of the test is ultimately interested in, namely, the way observed scores relate to true scores. More precisely, it can be proven that the correlation of observed scores on two parallel tests is equal to the squared correlation of observed scores (on either test) to true scores, i.e. $\rho_{OT}^2 = \rho_{OO'}$ (see Gulliksen 1950, 23;[19] Lord and Novick 1968, 3.1 and 3.3). This is useful, because parallel tests therefore allow us to talk about the way in which observed

[17] Lord and Novick (1968, ch. 3) define parallel tests in terms of equal true scores and equal variances of error.
[18] Together with other assumptions about the nature of error and true score, see discussions in section 2.4.1 above.
[19] Gulliksen expresses the result as the equality of the correlation of the observed score and the true score on the one hand and the square root of the correlation of the observed scores on parallel tests on the other.

scores on a test relate to the true scores – a relation that is in principle unobservable but valuable for understanding how faithfully (reliably!) observed scores track the true score.

All this has been said while bracketing the practically pressing issue of obtaining parallel tests *in practice*. We have thus far assumed that we have parallel tests and looked at what work they can be put to in estimating reliability. But clearly, since parallel tests are defined in terms of equivalent true scores, and true scores are unknown, it is difficult to identify parallel tests. Gulliksen argues that the following are "objective quantitative criteria" for identifying parallel tests: approximately equal means, equal standard deviations and equal intercorrelations (Gulliksen 1950, 14). However, not only is it difficult to find parallel tests (using these or other criteria proposed in the literature) it is also time and effort consuming to construct and administer two measures instead of just one (Cronbach 1951). The popular solution that emerged was to split the proposed test in half and estimate reliability in terms of the correlation between the two halves of the same test. That correlation came to be denoted as the *split-half coefficient*.

However, estimates of reliability end up different depending on how one splits the test (see references in Cronbach 1951). In other words, a test has multiple reliabilities, each corresponding to a different split of the test. Historically, this was considered a great shortcoming of the split-half approach. Several solutions were proposed in the psychometric literature (Tryon 1957) and in fact the theoretical literature is filled with different proposals for supposedly useful coefficients for estimation of reliability (Fiske 1971, ch. 8). The best known and by far the most frequently used is the *coefficient alpha* (Cortina 1993; Sijtsma 2009).[20] Building on earlier solutions to the split-half problem (in particular that presented by George Frederic Kuder and Marion Webster Richardson), Cronbach (1951) introduced the coefficient alpha and argued that alpha can be interpreted as the average of all possible split-half coefficients of a test. Coefficient alpha is today the most common measure of reliability, although it goes under a different name: *Cronbach's alpha* (Streiner 2003). The formula can be written as follows (see Cronbach 1951; Lord and Novick 1968, 89-90):[21]

[20] Sijtsma and Cortina provide contemporary overviews of the usage of and problems with Cronbach's alpha.

[21] So-called standardized alpha reads:

$$\alpha_{STD} = \frac{nr}{(1 + (1 - n)r)}$$

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum \sigma_{Y_i}^2}{\sigma_X^2}\right)$$

where *n* is the number of items,

$\sum \sigma_{Y_i}^2$ is the sum of the variances of scores on items $Y_1$, $Y_2$, … $Y_n$ in test X, and

$\sigma_X^2$ is the variance of the total test score on test X.

---

where n is the number of items, and
r is the average interitem correlation.

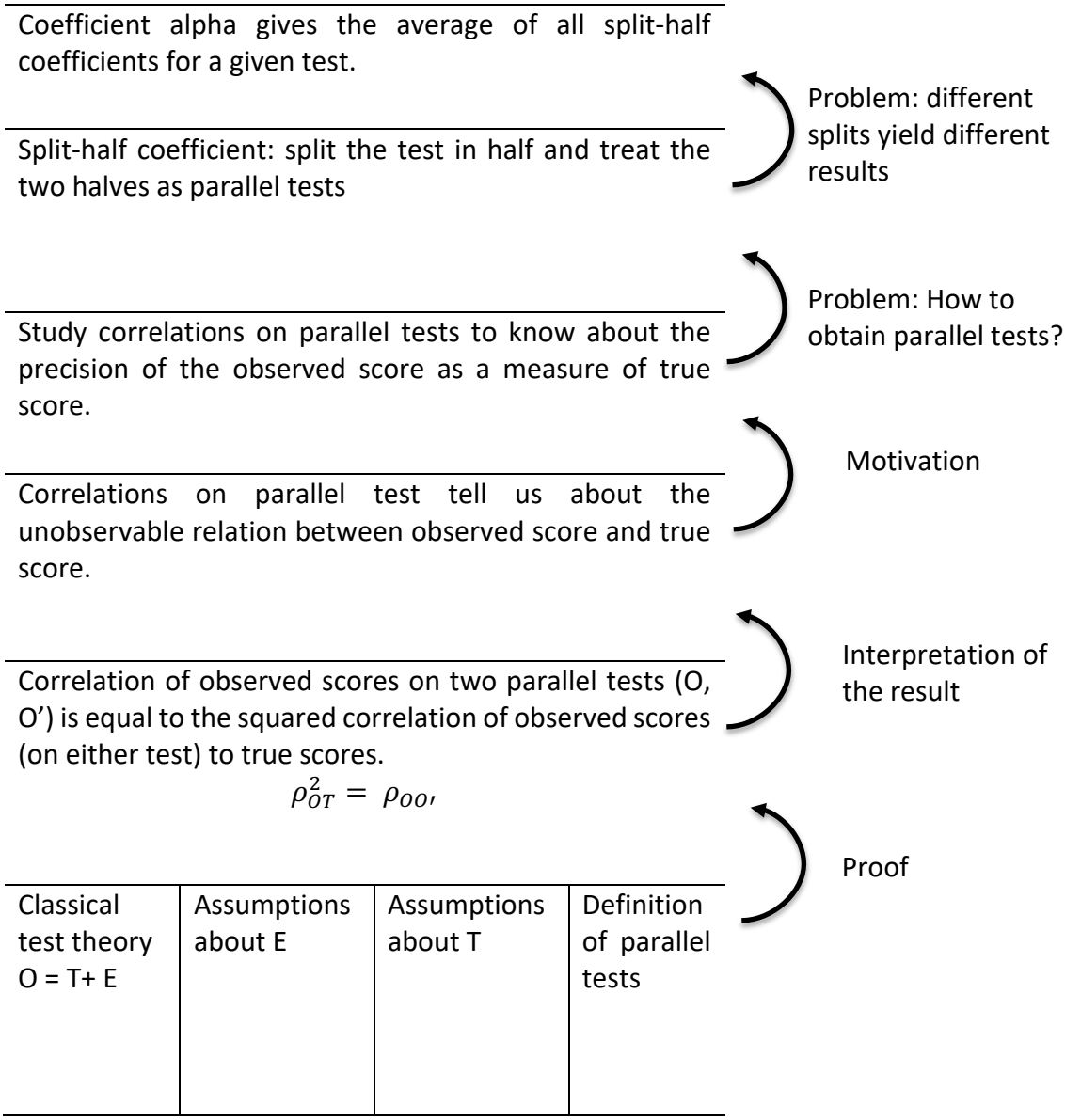


*Figure 4. From CTT to Cronbach's alpha. The arrows indicate the conceptual connection between two rows.*

We have now seen the most common measure of reliability (coefficient alpha) and investigated its background in classical test theory. Figure 4 summarizes these themes.[22] With these descriptions of the theoretical background of reliability estimation and the most commonly used coefficient of reliability, we are ready to understand some of the main controversies surrounding theories and tests of reliability.

### *2.5.3 Critique I: Is parallel tests reliability conceptually confused?*

Loevinger (1957, 676) argues that Gulliksen's notion of reliability is viciously circular:

> *Thus [in Gulliksen's 1950 book] parallel tests are defined in terms of their correlation (reliability), and reliability is defined as the correlation between parallel tests. It is not surprising that a theory which begins with a circularity ends in paradox.*

In view of my earlier summary, this criticism might strike the reader as odd. According to my review, Gulliksen does not define reliability and parallel tests in a directly circular manner. Rather, he defines reliability in terms of the correlation of the *observed scores* on parallel tests, and parallel tests in terms of equal (and thus perfectly correlated) *true scores*. Arguably, though, the problem of circularity does seep in if, in the absence of knowledge about true scores, one uses convergence on observed scores as a justification of the claim that the two tests are parallel. This is a potential way to object to Gulliksen, given that his proposed checks of whether two tests are parallel involve precisely the checking of observed similarities.

Loevinger's objection, however, has a further grounding in Gulliksen's work. In his chapter 3, which I have not discussed in this review, Gulliksen defines parallel tests "in terms of observable characteristics" (Gulliksen 1950, 28), namely, equal observed means and standard deviations, as well as equal intercorrelations between various

[22] The chain of inference I have presented here is a reconstruction of what I take to be the line of reasoning present in the reviewed literature. The adequacy of that chain of inference can of course be questioned from an argumentative perspective. My aim here has not been to evaluate but to describe, but for a detailed evaluation of these themes, see Borsboom 2005, ch. 2.

(proposed) parallel tests. To then marry this definition of parallel tests with the notion of reliability as (observed) correlation between parallel tests (Gulliksen 1950, 31, fn. 1), is, according to Loevinger, circular. Whether or not one agrees with Loevinger's conclusion, their exchange shows that even within a single author's work, there can be multiple notions of parallel tests at work. In other works, parallel tests are defined in terms of characteristics that differ from both of Gulliksen's proposed characterizations (e.g. Nunnally and Bernstein 1994, 223).

A final, empirical criticism of Gulliksen's notion of parallel tests is that finding genuinely parallel tests is difficult if not impossible. Embretson and Reise (2000, 21) make this point, arguing that Gulliksen's proposed "objective quantitative criteria" are rarely fulfilled. Moreover, with at least some specifications of the notion of parallel tests, it appears that there is no direct way of testing whether two actual tests in fact are parallel. Nunnally and Bernstein (1994, 226) argue as much, and in light of our previous discussion, it seems evident that at least one of Gulliksen's notions of parallel tests suffers from the same problem (see also Tryon 1957). All this indicates that there is no shared approach to the definition, interpretation and application of the notion of parallel tests.

### *2.5.4 Critique II: What's true about true score?*

Since the notion of reliability is based on CTT, problems with CTT manifest also as problems in reliability estimation. Recall that I introduced two definitions of true score in the sections on CTT. On the one hand, true score has been taken to refer to true ability, that is, the degree of the target attribute the test subject possesses (e.g. Gulliksen 1950, 4). In other applications, on the other hand, true score has been defined as "the average of all scores a person would receive if he or she took the test an infinite number of times" (Streiner 2003, 99). Thus, there are two prominent alternatives, which I have called the "true ability" interpretation of true score and the "infinite administrations" notion of true score.

Recalling now that reliability is defined in terms of correlations between parallel tests, and that this correlation has been shown to be equal to the squared correlation between true score and observed score, the relevance of different notions of the true score for interpretations of reliability coefficients becomes clear. Under different

notions of the true score, reliability is an indicator of different kinds of relationships: (squared) correlation of observed score and true ability or (squared) correlation of observed score and expected score on infinite administrations of the test.

We can explore the consequences of different interpretations of the true score a little further. On the "true ability approach", it is most intuitive to think of reliability coefficients as estimates of accuracy, i.e. the closeness of the test score to the real degree of the attribute (see Tal 2017, sec. 8.3 on this kind of notion of accuracy in general measurement theory). On the "infinite administrations approach", by contrast, reliability is better described as an indicator of repeatability or representativeness, because here the convergence of the observed score with the true score means that the observed score is close to, or representative of, results on multiple equivalent or similar tests.[23] In the latter case, a test instrument can be reliable but "wrong" about an attribute. Modifying an example from Kline (1998, 26), consider two 30 cm rulers, where each 1 cm division is wrong relative to the standard unit and slightly off relative to each other. The results from the two rulers on multiple measurements will correlate, and thus be reliable in the infinite administrations sense. But we would not call the results accurate, because the rulers are consistently wrong about the attribute of interest. This should already evoke the idea that the choice of the notion of true score has wide ranging consequences for what can justifiably be claimed based on tests of reliability.

As a final point, let me confirm what the reader might already expect: in applications, these two interpretations of the true score get confused and equated. Thus, we find, for example Andre De Champlain, the director of the *Medical Council of Canada*, writing in a manner that suggests such equivocation:

> *The candidate's true score, T, is defined as the expected value of the observed score over an infinite number of repeat administrations with the same examination. A true score can be thought of as the score that would be obtained*

[23] Note that while I am here using the same terms that occurred in some of the definitions of reliability in the beginning of this section, I do not mean to imply that the authors' choice of wording means they adopted this or that interpretation of true scores.

*if the examination was perfectly measuring the ability of interest (i.e. with no measurement error).* (De Champlain 2010, 109)

### *2.5.5 Critique III: Alpha's reign*

I mentioned above that there are multiple indices of reliability besides coefficient alpha (also known as Cronbach's alpha), but that coefficient alpha is by far the most popular and best-known in psychometric practice. Nonetheless, in the theoretical parts of psychometric literature, the interpretation and usefulness of coefficient alpha have been debated since its inception (Cortina 1993; Sijtsma 2009).[24]

One of the controversies around coefficient alpha concerns its interpretation, that is, what exactly a high or low coefficient alpha tells about the test of interest.[25] On one interpretation, alpha pertains to the *homogeneity* of the test roughly in the sense of "do all items measure a single dimension". On another interpretation, alpha pertains to the *internal consistency* of the test roughly in the sense of "are all items interrelated". As an example of how homogeneity and internal consistency come apart, one can contrast a test where all items measure a single aspect of depression, say, insomnia (homogeneity) and another test where different items measure different aspects of depression, e.g. insomnia, mood and guilt, but are all nonetheless related in the sense that they all pertain to depression (internal consistency). What complicates the interpretational debate considerably is that even the explications of homogeneity and internal consistency are debated (Lord and Novick 1968, 4.5; Sijtsma 2009).

Many authors argue that while coefficient alpha has been associated with internal consistency and homogeneity (under various explications of these properties), alpha is a poor measure of both properties (e.g. Sjitsma 2009). As a consequence, these authors propose that alpha's reign is unwarranted, and that psychometricians should pay much more attention to other reliability indices, such as *beta, omega* and *glb* (Sijtsma 2009; Revelle and Zinbarg 2009). The nature of these indices is not of importance here, but rather

[24] In his (2004), Lee Cronbach (with editorial assistance from R. J. Shavelson) argues that coefficient alpha is only a small component of a larger system of adequate reliability analysis. As Cronbach notes, this contrasts with the fact that alpha is so ubiquitously used in contemporary psychometrics.

[25] Coefficient alpha is also interpreted as a measure of the lower bound of reliability, when the requirement of parallelism is relaxed. See Borsboom (2005).

the observation that while alpha reigns in practice, the justification for this practice is debated.

### *2.5.6 What reliability to rely on?*

We have seen that although there are some more or less shared components to psychometricians treatments of reliability (such as the ubiquitous usage of coefficient alpha and the adoption of CTT as a background theory), interpretations and applications diverge radically from context to context. There is disagreement over the meaning of reliability (accuracy? repeatability?), definition of true scores (true ability? average on infinite tests?), definition and testing of parallel tests (same true scores? similar observed characteristics? which ones? how to test them?) and the meaning and usefulness of coefficient alpha. Table 5 summarizes these controversies.

Given these divergences in the details, even the seemingly shared components turn out to be only nominally shared. CTT might be the shared background theory, but depending on how one interprets true scores, the central equivalence of CTT amounts to a very different kind of theory. Furthermore, depending on how true scores and parallel tests are interpreted, very different conclusions are derived from reliability coefficients. Even though coefficient alpha is the go-to-coefficient, its meaning depends heavily on whether it is considered to inform about homogeneity, internal consistency or something else. Given all the divergence and controversy, it is best to say that there is no unified psychometric approach to reliability.

| Concept | Controversy |
|---|---|
| *Parallel test.* A concept used to explicate reliability in terms of correlations on two tests. An actual test with specific properties. | Empirical issue: how to establish that two tests are parallel. Conceptual issue: how to define reliability and parallel tests without circularity. |
| *True score.* A component of the classical test theory, stating that the observed score is a function of true score and random error. | True score has multiple meanings, which arguably lead to different conceptions of reliability. Problems of equivocation. |
| *Cronbach's alpha.* The most common index of reliability, also known as coefficient alpha. | Many alternative indices of reliability exist, some of which are arguably better suited for some contexts in which Cronbach's alpha is used. |

*Table 5. Controversial concepts in the reliability literature.*

### 2.6 No Capital V Validity

According to Kline (1998, 34), the standard, textbook definition of validity is that "[a] test is said to be valid if it measures what it purports to measure". This is, indeed, what we find in several books of psychometrics as well as in evaluations of psychometric theory. But the study of validity divides into numerous subcategories, that is, aspects of validity. Thus, there are textbooks making prescriptions of the *kinds* of validities that need to be established, and recommendations for doing so. These aspects of validity have come to be known with such names as *content validity, face validity, criterion validity, concurrent validity, predictive validity, structural validity, external validity, internal validity, discriminant validity, convergent validity* as well as *ecological validity, synthetic validity, incremental validity, metric validity, trait validity, nomological validity, practical validity*... And so on and so forth (cf. Markus and Borsboom 2013, sec. 1.2.2). In the following I will introduce some of the best-known types of validities in more detail and evaluate whether there is an overarching approach to validity within psychometrics.

#### *2.6.1 Criterion, predictive, and content validity*

*Criterion validity* is one of the oldest and most frequently occurring aspects of validity within the psychometric literature (Sireci 1998). The core idea is that to establish the validity of a measurement instrument, we need to check that it correlates highly with another measure, known as the criterion (e.g. Fiske 1971, 166). For example, one might argue that a novel measure of depression needs to correlate highly with the judgment of

the psychiatrists that have extensive experience with the treatment of depression – the psychiatrists' judgment is the criterion against which the measure is validated.

The obvious difficulty is the choice of criterion: what measure counts as a relevant criterion, that is, the measure with which the novel measure should correlate (e.g. Sireci 1998)?[26] Many have argued that what is required is that the criterion in some sense tests a similar thing as the novel measure is intended to measure – the example of expert judgment of depression illustrates this line of thinking (cf. Fiske 1971, 166; Kline 1998). Gulliksen (1950), by contrast, argues that the criterion needs to be chosen according to the intended purpose of the novel measure, in particular, the predictive role the novel test is meant to serve. Consequently, a test has as many criteria as it has predictive functions, and as many validities as it has criteria. For example, a given psychological test might have "one validity for predicting grades in English and a different for predicting grades in Latin" (ibid., 88). While Gulliksen (1950) calls this criterion validity, the emphasis on prediction locates him closer to the advocates of the importance of *predictive validity*.

Predictive validity is the idea that a measure is valid if it demonstrates utility in predicting future behaviour (Kline 1998, 36-37; Sireci 1998, 106-107). While the idea is simple, many problems have been noted. A major problem concerns the validity of the criterion (Fiske 1971, 167; Nunnally and Bernstein 1994, 97; cf. Sireci 1998). Fiske (1971, 167) provides an example. During World War II, psychologists developed tests that were intended to predict the aptitude for being a bombardier. As a criterion of validity, they used the candidate's accuracy on practice bombarding runs, given that that was much more practical than using actual combat situations as a criterion. It turned out that the criterion was unstable so that the accuracy on practice runs would vary greatly from day to day, across weather conditions, and so on. Thus, the criterion was essentially invalid, which evidently undermines the predictive validity of the proposed aptitude test.

Loevinger (1957) verbalizes another problem with predictive validity. According to her, prediction is a wrongheaded concern in psychometrics. In her view, the

[26] Reference to predictive and criterion validity has been particularly prominent in the context of achievement and ability testing, rather than the testing of, say, aspects of personality. The main reason for this is, arguably, that it is *prima facie* easier to come up with a criterion (predictive or otherwise) in the former context. Kline (1998) gives the example that criterions for an intelligence test are easy to identify among other achievement and ability measures, whereas it is less obvious what would be a valid and relevant criterion for extraversion.

proper aim of validation is to ensure that the measure captures a real attribute, not that it predicts some other test performance or behaviour. Accordingly, we find Loevinger writing (about herself) (1957, 640):

> *The writer believes that the most fruitful direction for the development of psychometric devices, and hence of psychometric theory, is toward measurement of traits which have real existence in some sense [and] that this orientation is antithetical to one which places first emphasis on prediction.*

Moving on to another kind of validity. Sireci (1998) maps out the history of *content validity*. While content validity has been a controversial concept since its inception (Sireci refers to Messick as a significant critic), it too gets frequently mentioned as important throughout the reviewed texts (Nunnally and Bernstein 1994; Rust and Golombok 2009; de Vet et al. 2011).[27] Broadly speaking, content validity tends to refer to the "adequacy with which a specified domain of content is sampled" (Nunnally and Bernstein 1994, 101; cf. Sireci 1998 for a more detailed analysis). Content validity is easiest understood (and most often applied) in the context of tests examining the degree to which a subject has mastered the content of a course or subject. For example, the content validity of a final exam of a psychometrics course should adequately cover the full domain of the course, i.e. the various subtopics, for the test to be content valid.

The term content validity has several meanings, which is why I chose to write "tends to refer to" above. Fiske (1971) uses content validity as a synonym of *face validity*, which stands for an expert judgment of "what a test measures". On this notion, content validity is the assessment of validity at face value (or appearance of validity), not the evaluation of the domain of applicability of the test. While Fiske (1971, 164) rejects content validity *qua* face validity, saying that it is "unwise to rely" on it, several contemporary psychometrics books list face validity as a significant aspect of psychometric validation. Face validity is sometimes treated as a component of the broader notion of content validity (de Vet et al. 2011) while at other times it is an aspect of validity on its own (Rust and

[27] In addition, Sireci (1998) argues that while Gulliksen (1950) does not mention content validity, his work contributed positively to the inception of this type of validity.

Golombok 2009, 78). Notably, even those who define content validity in terms of the domain coverage of the test rather than in terms of appearance of validity, and defend the necessity of content validity in this sense, do often argue for appeal to expert judgment as the determinant of content validity. But Sireci (1998) also thinks that statistical evidence can be employed for establishing content validity, while de Vet et al. (2011) argue that in many contexts the test takers themselves are appropriate evaluators of content validity.

Loevinger (1957) conceptualizes content validity in terms of domain coverage, and argues against the method by which content validity is usually assessed, namely, expert judgment. She argues that content validity, like predictive validity is “ad hoc” (ibid., 636), and that the judgment on content validity tends to change from investigator to investigator, even when the content validity of the same test is at issue. Loevinger's disappointment with predictive, criterion and content validity leads her to argue that all scientifically interesting validity is *construct validity*: “construct validity is the whole of the subject [of validation] from a systematic, scientific point of view” (ibid., 641).

### *2.6.2 Construct validity*

The concept of construct validity grew from the work of the *Committee on Psychological Tests* (1950-1954), which was established by the *American Psychological Association* (APA) to set up standards for adequate tests of measure validity. The work of the committee led to a report entitled *Technical recommendations for psychological tests and diagnostic techniques* (commonly referred to just as "Technical Recommendations"), in which the Committee's “chief innovation”, construct validity, was explicated. While the reports of the committee laid APA's official notion of construct validity, Cronbach and Meehl's 1955 “Construct Validity in Psychological Tests” is nowadays regarded as the classic statement of the notion of construct validity.

According to Cronbach and Meehl (1955, 282), “[c]onstruct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not ‘operationally defined.’” To carry out an assessment of construct validity, the construct must be theorized as a part of a *nomological network*, i.e. a network of statements connecting the target construct to other constructs by means of definitions and statements of laws. In particular, the nomological network must contain statements about

observable relations that manifest if the proposed measure captures the correct target construct. For example, we could have good grounds for theorizing the relationship between well-being and mental health such that observed results on an adequate measure of well-being should correlate with observed results on a measure of mental health. The validation occurs by testing the predicted observable relations: if they receive confirmation, that is an argument in favour of the test capturing the correct construct; if the relations are disconfirmed, that means that either the test is not appropriate or the nomological network is faulty (or both).[28]

What does the concept of construct validity mean for testing? Cronbach and Meehl suggest that much of the test practice that preceded the invention of construct validity can be usefully employed within the novel framework:

> *Many types of evidence are relevant to construct validity, including content validity, interitem correlations, intertest correlations, test-"criterion" correlations, studies of stability over time, and stability under experimental intervention. High correlations and high stability may constitute either favorable or unfavorable evidence for the proposed interpretation, depending on the theory surrounding the construct.* (Cronbach and Meehl 1955, 300)

While Cronbach and Meehl seem to allow almost anything to count as testing of construct validity, as long as the test results are interpreted in light of a nomological network, the term construct validity has later come to stand for a more specific testing practice. In particular, many theorists have emphasized that construct validation involves tests of *convergent* and *discriminant validity*, i.e. tests of how closely the proposed measure correlates with related/similar measures, and tests of the extent to which results on the proposed measure diverge from results on unrelated/dissimilar measures (D. T. Campbell and Fiske 1959; Nunnally and Bernstein 1994).

Cronbach and Meehl emphasize that construct validity is an approach intended for cases where the target construct is not defined by the operations that are meant to

[28] Nomological network was of course a concept logical empiricists developed and utilized, see Feigl (1950). Cronbach and Meehl reference the work of several logical empiricists in their (1955) paper.

measure it.[29] Loevinger (1957) also takes this to be a central feature of construct validation: she adamantly insists that construct validation concerns tests that aim to capture a real attribute rather than something operationally defined. She writes (642):

> *Construct connotes construction and artifice; yet what is at issue (in construct validation) is validity with respect to exactly what the psychologist does not construct: the validity of the test as a measure of traits which exist prior to and independently of the psychologist's act of measuring.*

In Loevinger's psychometric theory, construct validity divides into three mandatory and comprehensive components (1957, 686). They are as follows:

- *Substantive validity*: "The substantive component of validity is the extent to which the content of the items included in (and excluded from?) the test can be accounted for in terms of the trait believed to be measured and the context of measurement." (ibid., 661) For example, if one is interested in mathematical ability, items in the test should be selected so that they not only pertain to mathematical ability but also cover the full domain of mathematical ability (e.g. algebra, number theory, geometry). The relevant domain is defined by one's theory of mathematical ability and context of testing.
- *Structural validity*, which "refers to the extent to which structural relations between test items parallel the structural relations of other manifestations of the trait being measured." (ibid., 661) For example, two ability test items measuring mathematical ability should exhibit diverging response patterns, if one item is very difficult and the other is very easy. Such test response patterns parallel a theoretically expected pattern of behavioural manifestations, namely, that people with higher ability tend to pass items lower ability candidates do not pass.

[29] Operationally defined constructs will be explored in chapter 6.

- *External validity*, which involves comparisons of relationships between the proposed measure and other measures and indicators. For example, does a test of mathematical ability correlate with other measures that purportedly measure that same ability?

In his contributions, Messick (1989, 1995) diverges from Loevinger's approach in at least two ways. Firstly, unlike Loevinger, Messick does not focus on the realness of the attributes as the grounding and main concern of the construct validation exercise. Validity is not about whether the test captures something real – it is about the adequacy and usefulness of the inferences and actions that are made on the basis of the test scores. He writes (1995, 741; references in original):

> *Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment (Messick, 1989). Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. […] In particular, what needs to be valid is the meaning or interpretation of the score; as well as any implications for action that this meaning entails (Cronbach, 1971).*

Like Loevinger, Messick is critical of the traditional conception of validity, which includes content, criterion and construct validity. He too proposes a framework in which construct validity is the overall, main validity. Unlike Loevinger though – and this is the second difference between Messick's and Loevinger's respective notions – Messick divides construct validity into six aspects, none of which are absolutely mandatory. "What is required is a compelling argument that the available evidence justifies the test interpretation and use, even though some pertinent evidence had to be forgone." (Messick 1995, 744)

The six aspects that Messick proposes are meant to cover what content and criterion validity used to cover in the traditional conception, and add to them. Briefly stated, the six aspects are:

- *Content aspect* includes evidence that the test content is relevant for and representative of the construct of interest.
- *Substantive aspect* concerns theories and models of the processes that bring about the observed test behaviour, and evidence that the hypothesized processes in fact underlie the behaviour.
- *Structural aspect* reflects how well the scoring structure of the test reflects the structure of the construct domain of interest.
- *Generalization* concerns the extent to which the score interpretations generalize to other groups, tasks and circumstances.
- *External aspect* concerns evidence of convergent and discriminant validity.
- *Consequential aspect* is assessed when we "inquire whether the potential and actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but also at the same time consistent with other social values." (Messick 1995, 744)

### *2.6.3 Queen of validities?*

We have seen that there are many notions of validity at play in psychometric practice. The most notable ones are criterion validity, predictive validity, content validity and construct validity. We have also seen that even when authors use the same terms, often the concept of validity that term is thought to stand for is explicated in slightly different ways by different authors. The question then is whether there is any common, shared core that could be identified as the psychometric theory of validity.

Many commentators argue that there is more or less a consensus on construct validation as the overarching, main type of validity (Strauss and Smith 2009). Construct validity is also the aspect of psychometric theory that many philosophical commentators take as their target of analysis, as the core of psychometrics (Angner 2011; McClimans 2013; Alexandrova and Haybron 2016; also McClimans, Cano and Browne 2017 to an extent). And certainly, many of the texts I have reviewed do present construct validation as an improvement on weaker notions of validity, such as criterion and content validity (e.g. Fiske 1971, 164-167; also Kline 1998, 35). The fame of Cronbach and Meehl's

paper on construct validation also suggests that there is something special about construct validity. Perhaps there is, then, one core notion of psychometric validity, which is the proper unit of analysis when evaluating psychometrics in general?

The hype around construct validation seems indeed to indicate agreement among theoreticians. More specifically, they appear to agree that construct validity *qua* theory-referenced validation has something special going for it. In other words, they agree that statistical tests of the properties of a measure, e.g. its reliability, internal consistency, content, relations to other measures and so on, must be interpreted in terms of theoretical expectations of how a measure of the relevant attribute should behave. Despite the agreement, I would like to draw attention to three challenges to the claim that construct validity is the Capital V Validity that unifies all psychometricians.

Firstly, many of the theoreticians themselves present a caveat to the reign of construct validity. Cronbach and Meehl (1955, 282) write that: "construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined.'" (We have seen that Jane Loevinger also emphasizes the intimate link between construct validation and the realness of the attribute of interest.) The importance of construct validation is taken to depend on the aims and interests of the researcher: if the researcher's interest is prediction of a criterion or test performance in and of itself, then the appropriate concern is not construct validity but rather predictive/criterion validity and content validity, respectively. The relevance of construct validation is limited (at least in Cronbach and Meehl's classic exposition) to situations where the non-operationally defined attribute underlying the measure is of interest. Even for theoreticians, then, construct validity is not Capital V Validity across contexts, but only in contexts with specific aims (although this might be the most common aim).

Second, and more importantly, the theoreticians' constructions of construct validity underdetermine what should happen in validation practice. This is evident, for example, from Messick's and Cronbach and Meehl's lists (described above) of various considerations that *might* bear on construct validity, depending on situation. These lists are lax about the *exact* requirements of construct validity, beyond the requirement to interpret results in terms of a theory. Therefore, it is no wonder that many contemporary textbooks have opted for a narrower, easier-to-operationalize notion of construct

validation, which typically stresses convergences and divergences between measures. This narrower construct validation is often presented as complementary, rather than as superior to other validities. For example, Nunnally and Bernstein's influential textbook (1994) discusses predictive, criterion and construct validity, and treats them as complementary to one another. Rust and Golombok's 2009 textbook introduces altogether six types of validities to the student of psychometrics, only one of which is construct validity. De Vet et al. (2011) introduce the student of medical measurement to content validity, face validity, criterion validity as well as construct validity. They go as far as to state that (2013, 169) "[c]onstruct validation is often considered to be less powerful than criterion validation." In textbooks, then, construct validity is not the Capital V Validity.

Third, the narrower, easier-to-operationalize notion of construct validity seems to have caught on better in practice than the overarching, theory-focused notion. This can be seen from the fact that construct validity is usually reported as one among many validities, rather than as a section of its own where individual results pertaining to content, internal consistency, structure and so on are discussed in terms of a theory.[30] Thus construct validity appears to not be the Capital V Validity for practicing psychometricians. In fact, when we look at statistics pertaining to validation practice, it turns out that tests that count as construct validation (even under the narrower definition) do not get as much attention as other types of validities. Consider Anthoine et al.'s (2011) study of the validation of Patient Rated Outcome Measures, i.e. PROMs. It is notable, for example, that although discriminant validity is considered a key aspect of construct validity in many accounts (Messick 1995, Nunnally and Bernstein 1994; Campbell and Fiske 1959), Anthoine et al. (2011) find that only 17,5% of the PROM validation studies in their sample examined discriminant validity. By contrast, more than 90% of the examined studies looked at content validity *qua* "ability of an instrument to reflect the domain of interest and the conceptual definition of a construct". Surprisingly many (65,8%) also assess face validity *qua* "the ability of an instrument to be understandable" to the targeted population. Construct validation is clearly not the queen in practice.

[30] See e.g. Pavot and Diener (1993) on the SWLS; Reynolds and Kobak (1995) on HAM-D; García-Batista et al. (2018) on Beck Depression Inventory; Snyder et al. (2000) on various measures of depression.

Summing up, it appears to me that while theoretical literature provides reasons to think that construct validity is something like the Capital V Validity, validation practice seems not to concur with this idea. In more general terms, since construct validity has been described in many ways, the notion can be employed in arguments that have opposing conclusions and that still have claim to truth on the basis of some proportion of psychometric literature. In my view, construct validation – with its many meanings – is such a malleable concept that it is unlikely to be a fruitful methodological tool in the evaluation and improvement of psychometric measures. The reason is that such a malleable concept will easily lead interlocutors to talk past each other (e.g. some defending construct validation under the expansive notion, some critiquing it under the narrower one), and might give rise to practically unfalsifiable claims (e.g. by allowing one to raise different aspects of different readings of construct validity when responding to different criticisms).

Whether or not one agrees that construct validation is an unfruitful methodological concept for contemporary discussions, what is important for the purposes of this review is that almost across the board (maybe excluding Gulliksen 1950), validation tends to be construed as a multifaceted activity, which can be legitimately executed in different ways in different contexts (see Table 6 for a summary). Some authors cluster all (or a large portion of) these activities together and label their theory-referenced interpretation construct validation, others do not. The point is that there are many flavours of psychometric validation out there, and in this sense, there is no single, substantive Capital V Validity.

| | Construct validity | Predictive / Criterion validity | Content validity |
|---|---|---|---|
| **Gulliksen 1950** | Not considered. | Whole subject of validity. (88) | Not considered explicitly. (According to Sireci 1998, Gulliksen laid the grounds for what was later called content validity.) |
| **Cronbach and Meehl 1955** | Important when trait underlying the test behaviour is of central interest. E.g. palmar sweating as a test of anxiety proneness. | Important when criterion behaviour is of central interest. e.g. employer interested in predicting an employee's performance. | Important when test behaviour is of central interest. e.g. a work sample as a test (of proficiency in that work). |
| **Loevinger 1957** | "[W]hole of the subject" of validity.<br><br>Divides into three aspects. All aspects mandatory. | Ad hoc. Prediction is not the right concern – capturing a real trait is. | Ad hoc. Judgment of content validity tends to be arbitrary. But considerations over content relevant for construct validation. |
| **Fiske 1971** | Important improvement on criterion validity. | Practical value in some contexts, but often problematic. | "Unwise" to rely on content validity qua validity at face value, i.e. face validity*. |
| **Nunnally and Bernstein 1994** | Important but non-comprehensive.<br><br>Can be used to support predictive and content validity. | Important in some decision-making contexts | Important in achievement testing. |
| **Messick 1995** | Comprehensive main criterion of validity, pertaining to interpretation of the score.<br><br>Divides into six aspects. Some aspects can be left out. | Part of generalizability within construct validity.<br><br>May sometimes be omitted. | One of the six aspects of construct validity.<br><br>May sometimes be omitted. |
| **Rust and Golombok 2009** | "Primary form of validation" (80) | Important when prediction is the aim of testing | "Fundamental to psychometrics" (79)<br>Include face validity as a separate, non-trivial criterion*. |

*Table 6. Comparison of notions of validity. *Note the interplay between two different meanings of content validity.*

### 2.7 No Capital P Psychometrics

Validity, reliability and modelling are all major themes in psychometrics. Within each of these themes, there are methods and techniques that have different, sometimes conflicting, and frequently debated rationales and interpretations. Table 7 summarizes some of these conflicts.

| Theme | Question | Conflicting answers | Relevance |
|---|---|---|---|
| Models | Is CTT outdated? | Simplicity of the framework<br>vs.<br>shortcomings vis-à-vis IRT | Establishing the appropriate test practice |
| Models | What kind of IRT model-fitting practice is apt? | Discard model if data doesn't fit<br>vs.<br>Discard data/test if model doesn't fit | Establishing the appropriate IRT practice |
| Reliability | What is a true score? | True ability<br>vs.<br>the infinite administrations interpretation<br>vs.<br>other interpretations | Different interpretations lead to different notions of reliability |
| Reliability | Why is coefficient alpha measured? | To establish internal consistency<br>vs.<br>to establish homogeneity<br>vs.<br>because of the norms of the field | There are arguably better, less well-known indices of reliability |
| Validity | What is construct validity? | Theory-referenced interpretation of measure properties<br>vs.<br>Assessment of convergent and discriminant validity | Establishing the claims construct validity can justifiably vouch for. Establishing cross-study comparability |

*Table 7. Some conflicts within each psychometric theme.*

When we add to these conflicts the fact that there are also conflicts over relations between these themes, it becomes, I believe, plausible that there is no unified framework for psychometric testing. What does that mean, concretely? Two psychometricians might set out to validate the same measure, using the same data, and come up with the opposite answer on each pair of conclusions: valid/not valid, reliable/not

reliable, fits the appropriate model/does not fit the appropriate model. Nonetheless, both psychometricians might have done a flawless psychometric job. They may each have multiple defining psychometric works on their side as well as the support of contemporary textbooks and established conventions.

The messiness of psychometrics may or may not be surprising. It may or may not be a bad thing (some clearly think it is, see e.g. Suppes and Zinnes 1962, quoted in section 2.1). In my view, the messiness of psychometrics readily warrants at least two claims. One: *there is much work to be done in terms of establishing and promoting the legitimate uses of psychometric technique vis-à-vis different aims*. Two: *we should be careful not to overgeneralize our conclusions when analysing the functions psychometrics can legitimately serve*. Motivated by these two observations, I will now move on to formulate criteria in terms of which psychometric methods can be evaluated. The next chapter provides a criterion of measurement-relevant representation.

# 3. Representation Minimalism

## 3.1 Representation, not off the shelf

Measurement is representation. That much is usually agreed upon. But what kind of representation, and how does one attain it? In this chapter I set out to define minimal constraints on the kind of representation measurement requires. Accordingly, my definition is called *Representation Minimalism*. Representation Minimalism is not a definition of all that measurement requires. Rather, the ultimate purpose of the definition is to help evaluate the potency of psychometric validation techniques in terms of one necessary aspect of measurement, that is, their ability to generate appropriate representations of target systems.

Why propose a new definition, rather than pick one off the shelf – surely there is plenty to choose from, given that philosophers have talked about representation for thousands of years?[31] It is true that plenty of articles and books have been dedicated to the topic of representation in general, and representation in measurement in particular.[32] The best-known and most ambitious account of the kind of representation *measurement* requires is called the Representational Theory of Measurement (RTM).[33] RTM received its authorative expression in three volumes entitled *Foundations of Measurement*, which were written by philosopher Patrick Suppes and psychologists R. Duncan Luce, Amos Tversky and David Krantz and which appeared in 1971, 1989 and 1991. Since then, RTM has played a key role in decision-theory in economics and philosophy, as well as in economic theory more generally.

The problem with building directly on RTM is that the content and implications of RTM are debated. Some argue that RTM is a formal theory specifying conditions of measurement, others argue that RTM is an epistemological theory specifying how to go about measuring (Tal 2012 introduces alternative interpretations). Some argue that RTM is useful for measurement (Heilmann 2015; Vessonen 2017), others argue that it

[31] Some conceptions of representation (which are still debated today) have their roots in Plato's *The Republic*. See Frigg and Nguyen (2016) for an overview.

[32] For example Hacking (1983); Morgan and Morrison (1999); van Fraassen (2008).

[33] Others might prefer to call RTM an approach to measurement rather than an approach to measurement-relevant representation – I have done so myself in Vessonen (2017). The reason I call RTM an approach to measurement-relevant representation will become clear later in the chapter.

may be redundant in some contexts (Angner 2011). Some argue that RTM may have formal merits but that it is useless for the practical execution of measurement (Reiss 2008; Mari et al. 2017), still others say that RTM fails as an approach to measurement and should be replaced by something else (Michell 2005). Some read RTM as a realist approach to measurement, others view it as an anti-realist or operationalist approach. Due to these controversies, and the heavy, formal presentation of RTM, I find it unwise to build directly on the RTM basis in evaluations of psychometrics.

However, the benefit of a rich, critical literature on representation (as it is conceptualized in RTM) is that it signposts controversial aspects of the subject. The literature on RTM therefore provides a roadmap of pitfalls to avoid when developing an improved account of measurement-relevant representation. After outlining Representation Minimalism, I will use the literature on RTM to show that Representation Minimalism avoids many of the worries that scholars have had about the conception of representation that RTM *is taken to advocate*. I emphasised "*taken to advocate"*, because it is my view that RTM is often misunderstood by its critics. Indeed, I will argue that when RTM is appropriately interpreted, RTM and Representation Minimalism are allies, not alternatives.

The chapter proceeds as follows. In section 3.2, I explicate Representation Minimalism. In section 3.3, I show how Representation Minimalism connects with RTM. Section 3.4 explicates the relation between Representation Minimalism and the practice of measure validation. Section 3.5 discusses Representation Minimalism and the realism/operationalism distinction. Section 3.6 concludes.

## 3.2 Representation Minimalism

### *3.2.1 Scales in measurement practice*

Consider an assignment of numerals to types of minerals:

Quartz ← 7

Calcite ← 3

The assignment consists of completely useless uninterpreted symbols unless we know what the symbols are meant to represent or be informative of. The user of these numerals wants to know (at least) two things before she allows herself to interpret the

numerals as meaningful measurement results. First, what attribute do the numerals pertain to? Do they indicate the weight or surface temperature or hardness of two mineral samples, or something else entirely? Second: what measurement scale are we dealing with? This section will focus on this latter question. The notion of *scale* is a crucial tool for explicating measurement-relevant representation, which is why we need to familiarize ourselves with this concept.

Scales are part and parcel of the conceptual and statistical toolbox of measurement. Hence, scales are introduced in most textbooks and introductory courses on social scientific measurement (e.g. Lord and Novick 1968; Fiske 1971; Nunnally and Bernstein 1994; Kline 1998; Embretson and Reise 2000; Howell 2010; Osherson and Lane 2018). But more familiar attributes from physical sciences, weight, temperature and hardness, are simpler cases for illustrating the most common measurement scale types, which were originally distinguished by psychophysicist S. S. Stevens (1946).[34] The following paragraph provides (what I take to be) an uncontroversial, broad brushstrokes account of how scale types are usually treated.

Our familiarity with weight measurement tells us that if the numbers indicate weight measured in grams, we can say, for example, that the ratio of the quartz sample's weight to that of the calcite sample is 2.3. In measurement jargon, we say that such comparisons are meaningful because weight is measured on a ratio scale. By contrast, we know from everyday usage of temperature scales that it is not sensible to say that the ratio of the temperature of the quartz sample to the temperature of the calcite sample is 2,3, when both are measured in centigrade. We could, however, compare the ratio of the *difference* between the temperatures of the two samples to the difference between the temperatures of some other two samples. In the language of measurement theory, we say that such comparisons are meaningful because the measurements are on an interval scale. Finally, if the numerals are measurements of mineral hardness on the less familiar Mohs hardness scale, the only thing the numerals are informative of is ordering. Quartz is harder than calcite, we can say, knowing that it has been assigned a higher value. But based on

[34] There are alternative classifications of scale types. See e.g. Velleman and Wilkinson (1993).

measurements on Mohs hardness scale, we know nothing about how much harder quartz is. That is because Mohs hardness scale for minerals is an ordinal scale.

These crude characterizations should be uncontroversial and acceptable to most scientists and philosophers who deal with measurement scales. It is also widely accepted that different scale types tend to allow for different arithmetic and statistical operations, although there is disagreement on how strict one should be regarding these rules (Stevens 1951; Borgatta and Bohrnstedt 1980; Luce et al. 1990; Velleman and Wilkinson 1993). Many social scientists and statisticians agree that one cannot, for example, sensibly take the mean of ordinal values or meaningfully add ordinal values to each other. Another example: in empirical social sciences, undergraduates must learn by heart that ordinal variables require a Spearman correlation test while Pearson correlation test can only be applied to variables measured on an interval or a ratio scale, and so on for other tests. It must immediately be added that the "rules" for identifying permissible statistical tests based on scale type are constantly debated and even more often broken. But for now, it suffices to remember that scale types set some constraints on what arithmetic and statistical operations can be fruitfully applied to measurement results.

I think it is appropriate to say, based on the foregoing, that scales are typically treated as representational assumptions. For example, when they pertain to the attribute weight measured on a ratio scale, the numbers 7 and 3 (from the previous example) *represent* the empirically established fact that in some situation of interest the quartz sample is 4 grams heavier than the calcite sample and that the ratio of the quartz sample's weight to that of the calcite sample is 2.3. The following examples from introductory resources enforce that a representational reading of scales is common:

> *[When dealing with an ordinal measure of stress] we do not assume, for example, that the difference between 10 and 15 points* ***represents*** *the same difference in stress as the difference between 15 and 20 points. Distinctions of that sort must be left to interval scales.* (Howell 2010, 7 emphasis added).

> *Interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahrenheit scale of*

*temperature. The difference between 30 degrees and 40 degrees* ***represents*** *the same temperature difference as the difference between 80 degrees and 90 degrees.* (Osherson and Lane 2018, emphasis added)

The common reading of scales, then, is that claims about scale types are claims about representation – e.g. "Measure M represents attribute A on scale S". The next section capitalizes on this notion of scales to explicate the kind of representation measurement requires.

### *3.2.2 Representation Minimalism – a definition of measurement-relevant representation*

With a familiarity with common scale types, we are ready to attempt the explication of the kind of representation measurement requires. Consider the following definition, my first approximation of the kind of representation measurement requires, i.e. measurement-relevant representation.

**R1**. A numerical representation is appropriate when specified relations in the representing numerical system mirror[35] empirical relations between entities.

As with many philosophical theories of scientific representation, this definition of representation capitalizes on our intuitions of structural similarity. The need for structural similarity is captured in the requirement that *relations in the representing numerical system mirror empirical relations between entities.* In other words, the relations that exist between numbers that are assigned to entities need to be similar to the relations that exist between those entities. For example, *ordering* of numerals should mirror *ordering* of entities, *equalities* of numbers should mirror *equalities* in the degree to which two entities possess a property, and so on, where the italicized words designate what structural aspects of the numerical system and the empirical system are similar to each other.

[35] This definition could be formed using terms such as *reflects or maps* where all these verbs are meant to connote something similar.

Of course, a general definition of representation in terms of similarity is notoriously elusive, not least because "everything is similar to everything else" *in some respects* (Isaac 2013; cf. McLendon 1955). To avoid a trivial definition of representation, one must say something about *the kinds of similarities* that are relevant for representation. Fortunately, in the measurement context, common scale types allow us to enumerate some similarity relations that are useful for representation without recourse to a general definition. The most common, measurement-relevant mirrorings are ones where: i) order relations between numbers mirror order relations between entities (ordinal scales), ii) (in)equalities of differences between numbers mirror (in)equalities of differences between entities (interval scales), and iii) (in)equalities of ratios of numbers mirror (in)equalities of ratios of entities (ratio scales) when entities are considered in terms of some attribute. These similarity relations are useful in virtue of being intuitive and recognizable to most people who are involved in measurement activities and who are therefore familiar with scales.[36] There are other scale classifications and thus other interesting and potentially useful measurement-related mirrorings. For the purposes of the present discussion, the most common scale types suffice though.

The term "specified" is important in the definition: *A numerical representation is appropriate when specified relations in the representing numerical system mirror empirical relations between entities*. Its role is to signal that, while different kinds of similarity relations (e.g. ordering, inequality and equality of differences) can justifiably underwrite measurement-relevant representation, it is important to be explicit about the relations a particular numerical assignment mirrors. On their own, numbers lure us to apply familiar operations such as addition and calculation of averages. But in measurement, the results of these operations only have a meaning if the numbers have a specific mirroring relation to the entities that are of interest to the measurer. The average of numerals that signal mere ordering has no empirical interpretation, for example. Hence, for the numerical representation to be appropriate, one must be clear about what similarity relations are being represented. The relevant relations need to be *specified*.

[36] I would hypothesize that these relations are intuitive to most laypeople too, simply because laypeople are familiar with common measurement instruments such as tape measures and kitchen scales. I am not aware of any empirical research on this question.

Our definition, R1, is still incomplete. Recall how in section 3.2.1 we noted that a purportedly measurement-relevant numerical assignment raises (at least) two questions: one about scale and one about the attribute the numerical assignment pertains to. But our current definition of measurement-relevant representation says nothing about the attribute that characterizes the empirical relations that are numerically represented. In principle, in the absence of any mention of the relevant attribute, R1 might lead to the following situation. Consider the following information about three individuals, Rei, Jalo and Kailo, and their orderings in terms of height and weight:

| | |
|---|---|
| Height | Kailo $\succ$ Rei $\succ$ Jalo $\prec$ Kailo |
| Weight | Rei $\succ$ Kailo $\succ$ Jalo $\prec$ Rei |

With our current definition, the following numerical assignment might be said to be an appropriate, measurement-relevant representation: Rei $\leftarrow$ 2, Kailo $\leftarrow$ 3, Jalo $\leftarrow$ 1.[37] Such a numerical assignment would mirror the following order relations: Kailo is taller than Rei, Rei is heavier than Jalo, and Kailo is taller and heavier than Jalo. But the representation is odd and difficult to interpret. We would typically not consider such a representation appropriate.

I therefore propose the following amendment to my previous definition:

**R2**. In measurement, a numerical representation is appropriate when specified relations in the representing numerical system mirror empirical relations between entities, when entities are considered in terms of the target attribute.

To form R2, I have simply added "considered in terms of the target attribute" at the end of R1. What are target attributes and what is it for entities to be considered in terms of such an attribute? There is extensive discussion of attributes (or properties) in philosophical literature,[38] and we cannot delve into that literature here without losing sight

[37] The numerals designate orderings, such that a higher number designates the heavier/taller individual and an equal number designates equal height/weight.
[38] e.g. Galluzzo and Loux (2015); Marmodoro and Yates (2016).

of measurement. Instead I propose that at a minimum, most evaluators must regard relations between entities as recognizably similar in order for the relations to count as being "in terms of the target attribute". For example, it will not do to talk about measurement, if the representing numerical system mirrors relations between three objects *a, b* and *c* such that *a* and *b* are ordered in terms of (what is commonly identified as the attribute) weight, *b* and *c* are ordered in terms of (what is commonly identified as the attribute) height, and *a* and *c* are ordered in terms of (what is commonly identified as the attribute) surface temperature.

In addition (and trivial-soundingly, but this will become important), the target attribute must be the attribute or feature the user of the measurement instrument is *interested in* – hence the word "target". For example, it won't do to tell me to consider relations between entities using a ruler, if the attribute I am interested in is depression (and no advice is given on how to read the ruler in terms of depression rather than length). This is because as a measure-user, I am unlikely to accept the ruler as pertaining to the attribute I am interested in, i.e. depression.

Of course, sometimes it is up to debate as to whether two types of relational statements are recognizably in terms of the same attribute. For example: are depression and anxiety linked intimately enough to warrant treating them as a single attribute in terms of which people can be ordered? I think my definition is useful even if I steer clear from any strong criteria for demarcating attributes from each other. In controversial cases (that is, when it is not clear that relations in the target empirical system are "in terms of the target attribute") the resolution must come via argumentation and evidence of (non-)similarity. For example, in the case of depression and anxiety, one might appeal to laypeople's and experts' intuitions about depression, philosophical argumentation, arguments concerning the reality of an attribute (realism vs. anti-realism), evidence of differences and similarities in causal relations that depression and anxiety enter in, and so on and so forth. What is required, though, is consistency: for example, if depression is deemed distinct from anxiety, one must not use relations in terms of anxiety as the (sole) grounding for claims about depression.

Figure 5 summarizes the resulting notion of measurement-relevant representation, explaining its key components. The definition is minimalist in many ways,

and I will therefore call it Representation Minimalism (ReMi). What I mean by minimalism is that the definition sets minimal constraints on the required representation – this is what measurement requires at minimum, although it might (be argued to) require a lot more in addition. First, ReMi says nothing about the exact relations that need to be mirrored. What matters is that the relations are specified and consistently used, not that they are, for example, all and only the relations that interval and ratio scales are taken to represent. Hence ReMi is open to the usefulness of scale types that go beyond the classic division to ordinal, interval and ratio scales. Second, ReMi is minimalist regarding the kinds of attributes that are represented. More specifically, ReMi is not tied to a realist or any other metaphysical stance regarding attributes. It only requires that what is represented is what the measurer is interested in representing and that the represented relations are recognizably similar in the sense explicated above (i.e. ordering *a* and *b* in terms of weight, and *b* and *c* in terms of height will not do as an ordinal representation). Such minimalism does *not* mean that realists or proponents of other metaphysical views (concerning attributes) cannot commit to ReMi. ReMi sets minimal constraints of measurement-relevant representation, which the metaphysically-inclined may supplement as they see fit.

I believe that ReMi is a necessary condition for measurement, and the rest of this chapter is dedicated to defending ReMi as such. But ReMi is likely not a sufficient condition for measurement. A full-blown account of measurement would likely specify how entities and the measurement instrument (causally) interact in order to produce an adequate representation. For the purposes of evaluating psychometric validation techniques, and their ability to confirm measurement properties, it turns out that ReMi suffices, that is, a full-blown theory of measurement is not needed in that evaluation.

The rest of this chapter divides into three distinct sections, which all deal with different aspects of ReMi. These aspects are: the relation of ReMi to RTM (section 3.3), the relation of ReMi to validation of scale type assumptions (section 3.4), and the relation of ReMi to operationalist and realist notions of empirical relations (section 3.5).

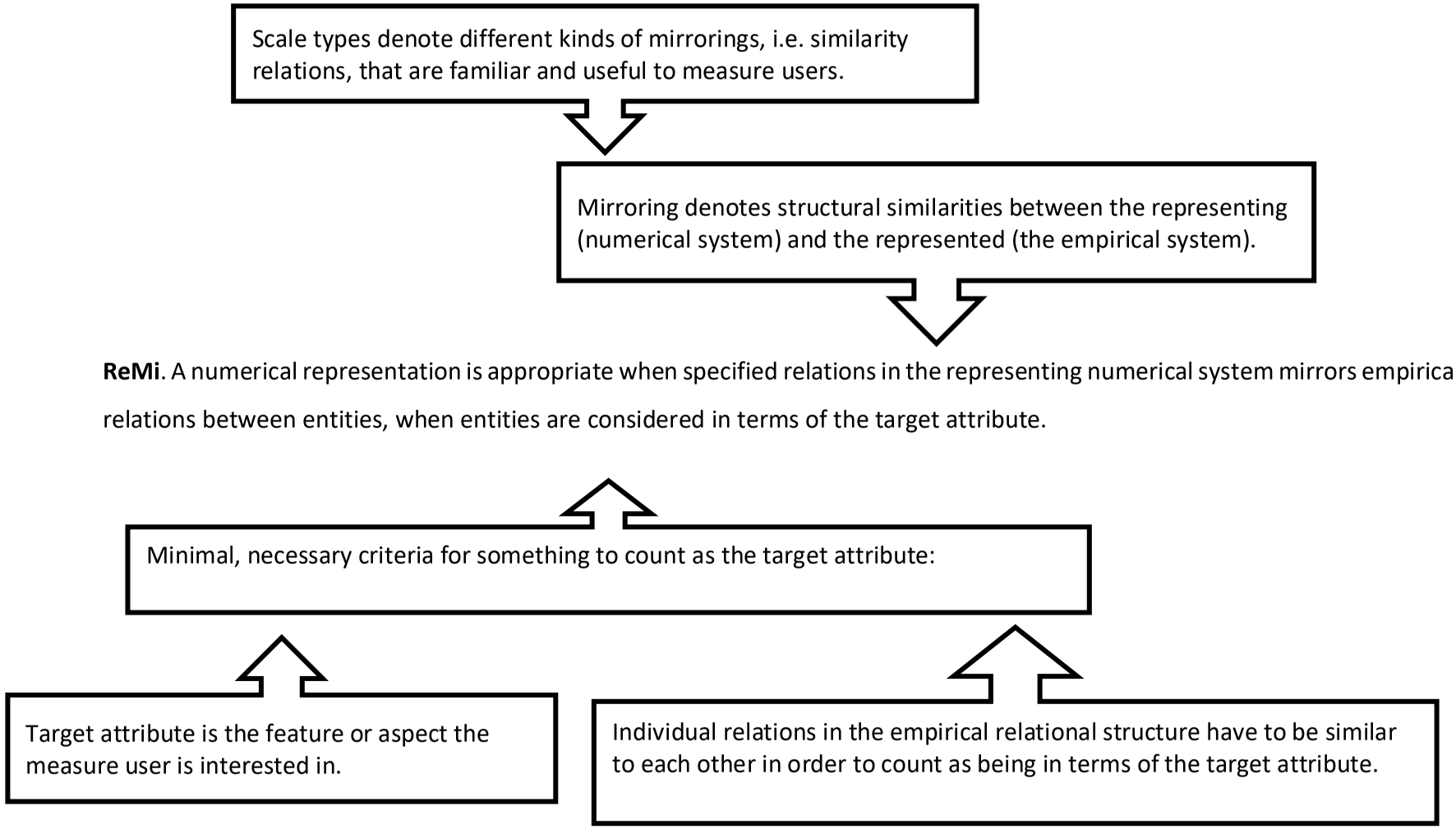


*Figure 5. A summary of Representation Minimalism.*

## 3.3 Formal foundations

### *3.3.1 Mathematics of scales*

The Representational Theory of Measurement (RTM) is the most ambitious attempt to theorize measurement-relevant representation. This section will outline the basic tenets of RTM, argue for a specific interpretation of RTM (since interpretations are debated), and show that under the defended interpretation, RTM is the formal foundation of Representation Minimalism.

According to the authors of *Foundations of Measurement*, where RTM received its canonical expression, measurement involves "the construction of homomorphisms [..] from empirical relational structures of interest to numerical structures that are useful." (Krantz et al. 1971, 9). In mathematics, *homomorphisms* are many-to-one mappings from sets to other sets. In RTM, these mappings are established between specific types of sets, that is, the mappings relate *empirical relational structures* to numerical (relational) structures. Foundational measurement theorists commonly distinguish between four types of homomorphisms: ratio, interval, ordinal and nominal. As one can guess, types of homomorphisms are in fact what we have previously called scales. The technical term "mapping" that proponents of RTM use serves the same role as the term "mirroring" serves in ReMi.

I have already said that informally, differences between scales can be thought of as differences in what information the numbers (numerals) represent about the targets of measurement. For example, if we are dealing with an ordinal scale, the numerals reflect or mirror the ordering of the entities while interval scales are informative of order *and* equality and inequality of differences between entities. But formally, different scales i.e. homomorphisms are characterized by the types of transformations they allow, that is, what are the rules for changing the numbers in a given numerical assignment. Ordinal scales, such as Mohs hardness scale for minerals, allow monotonic increasing transformations of the form $\varphi \rightarrow f(\varphi)$, where f is "a strictly increasing real-valued function of a real variable" (Krantz et al. 1971, 11). These transformations are permissible, because a rule-bindingly transformed numerical assignment continues to represent the same empirical relational system as the

original numerical assignment.[39] For example, if we are dealing with the ordinal Mohs hardness scale, the numerals in the numerical assignment

Quartz ← 7,

Calcite ← 3

could be transformed by adding 2 to both numbers or multiplying them by 2, and the resulting numerical assignments would continue to mirror the structure ordinal scales are informative of, namely, ordering relations between entities (i.e. quartz would continue to be assigned a higher value which corresponds to the fact that it is harder). For interval scales, e.g. temperature measured in degrees Celsius or degrees Fahrenheit, the permissible transformations are of the form $\varphi \rightarrow \alpha\varphi+b$, $\alpha>0$. Ratio scales, such as length and weight, allow for multiplicative transformation of the form $\varphi \rightarrow \alpha\varphi$, $\alpha>0$. These latter two scales are also called quantitative scales.

So far, I have largely re-described the content of the previous section in the more technical and rigorous language of RTM. The core idea of RTM that we have not yet explored is that in order for relations between entities to be measured on a specific scale, those empirical relations have to fulfil certain constraints. RTM states these constraints in axioms. For example, the axioms that pertain to an ordinal scale are:

Let A be a finite set of objects, and $\succcurlyeq$ a binary relation on A. The relational structure $(\succcurlyeq, A)$ can be meaningfully represented on an ordinal scale, iff for all $a, b, c \in A$,

1. Connectedness: Either $a \succcurlyeq b$ or $b \succcurlyeq a$, and
2. Transitivity: If $a \succcurlyeq b$ and $b \succcurlyeq c$, then $a \succcurlyeq c$.

For example: the set A of objects denotes a set of minerals, and the relation $\succcurlyeq$ denotes a relation in terms of hardness, i.e. $a \succcurlyeq b$ is interpreted as "a is at least as hard as b".

[39] Those versed in the technical literature might find the terms (and distinction between) "transformation rules" and "meaningfulness" helpful in this context. I shall not introduce this distinction here because I want to give an accessible gloss of RTM so that its connections to the practice of psychological measurement become apparent. A technical discussion of these concepts can be found in *Foundations of Measurement* (Krantz et al. 1971, Suppes et al. 1989, Luce et al. 1991).

Where do such constraints come from? The above claim about axioms of ordinal measurement, just like other axiomatizations in RTM, are backed up by theorems. In fact, proofs of two types of theorems fill the pages of *Foundations of Measurement* and other publications in the RTM tradition. A representation theorem establishes that if (sometimes *if and only if*) a given empirical relational structure of interest satisfies certain (non-contradictory) axioms, such as the ones described above, then a homomorphism φ to a certain numerical structure can be established. A uniqueness theorem establishes the permissible transformations of φ that also yield a homomorphism to the same numerical structure. In other words, the uniqueness theorem determines the scale type of the numerical assignment.

For example: if the hardness relation $\succcurlyeq$ satisfies connectedness and transitivity, then one can prove a representation theorem: there is a function φ from A to the set of real numbers such that for all minerals a and b in A, $a \succcurlyeq b$ iff $\varphi(a) \geq \varphi(b)$. In informal terms, what is proven is that the hardness relation $\succcurlyeq$ holds between a and b if and only if the number associated with a is greater than or equal to the number associated to b. Another function φ' has the same property and thus constitutes a homomorphism to the same numerical structure as φ iff there is a strictly increasing function f such that for all a in A, $\varphi'(a) = f[\varphi(a)]$. In informal terms, in this case φ' is a permissible transformation of φ as long as it preserves the order of the numbers assigned to the objects. The upshot of the proofs is that mineral hardness can be represented on an ordinal scale.

### *3.3.2 What is this thing called RTM?*

As noted in the introduction, multiple controversies surround RTM. Some contributors, for example, seem to think that RTM provides an account of how to go about measuring in practice, e.g. how to validate a measure or how to confirm scale type assumptions. More concretely, some scholars argue that RTM requires *direct observations* of relations that *perfectly* satisfy the relevant axioms (e.g. Michell 1986; Mari 2005; Borsboom 2005; Angner 2011).[40] From the observation that most successful measurement practices usually make no reference to direct observations or the fulfilment of axioms, it is then concluded that RTM is simply wrong about what successful measurement requires.

[40] Borsboom (2005, ch. 4) eventually ends up characterizing RTM as a rational reconstruction of the measurement process, therefore departing from the observation-tied interpretation he first assigns to RTM.

I think this critique of RTM does not hold water. This is because the critics have an incorrect and unfruitful interpretation of RTM. In my view, RTM is not a how-to account: it does not tell us anything about the kinds of observations that are needed to establish that a measurement scale is instantiated in a specific measurement context. RTM is merely a formal theory of what scale type assumptions entail (for defences of this reading of RTM, see Narens and Luce 1993; Decoene, Onghena, and Janssen 1995; Heilmann 2015; Vessonen 2017).

Now, it is tricky to distill the *correct* interpretation of RTM from the literature. This is because several people have contributed to laying the foundations of the approach, and there's no guarantee that the exact specification of the approach is constant across individuals or across an individual's career. There nonetheless seems to be ample textual evidence that the authors of *Foundations of Measurement* did not believe that RTM requires direct observational evidence of perfect fulfilment of axioms. Krantz and others write that "[t]he axioms purport to describe relations, perhaps idealized in some fashion, among certain *potential* observations" (Krantz et al. 1971, 26–27, italics added). Usually observations do not conform to the axiomatic conditions because of error: the experimental setting does not perfectly represent the relations in terms of the attribute we are interested in.[41] One possible solution according to the authors is to consider relational statements such as $a \succcurlyeq b$, not as statements about observations, but as theoretical statements inferred from the data (Suppes et al. 1989, 300). In fact, Suppes et al. (ibid.) write that this way of dealing with the errors that emerge in observational settings is "perhaps implicit in most of the axiomatic work on measurement e.g. Volume 1 [of *Foundations of Measurement*]".[42] This textual evidence suggests that the founders of RTM were not as observation-obsessed as the critics say they were.

We could of course continue the debate by rummaging through *Foundations of Measurement* (and other works) in search for evidence for this or that interpretation of what the founders of RTM really meant. But in my view our time is better spent thinking about what interpretation of RTM is most useful going forward, and then seeing how that interpretation

[41] Contrary to what some authors suggest, the authors of *Foundations of Measurement* were aware of the problem of error (Krantz et al. 1971, ch. 1) but they did not regard it as a problem that falls within the purview of measurement theory (Suppes et al. 1989, 300). It is therefore no wonder that error is not the focus of *Foundations of Measurement*.

[42] They also introduce other ways of dealing with such noisy data, such as random variable structures and structures where the relational constraints are expressed in probabilistic terms (Suppes et al. 1989 ch. 16).

helps us solve contemporary measurement problems. It seems that the observation-tied interpretation of RTM is useless, not least because the proponents of that view would need to draw a line between observable and unobservable relational structures in order to police illegitimate applications of RTM. The place of that dividing line is of course notoriously contested. Since many of the criticisms of RTM problematize precisely the alleged "how to" aspects of the theory, it seems plausible that our best epistemic bet is to interpret RTM as a formal theory of measurability (or numerical representability), not as a theory about how to go about measuring.[43] This is why I opt for the formal interpretation of RTM.

### *3.3.3 RTM and Representation Minimalism*

If RTM is treated as an account of the formal requirements of measurability, then RTM and Representation Minimalism are closely connected. In fact, RTM provides rigorous characterizations of the kinds of mirroring relations ReMi requires. ReMi relies on scale types to enumerate measurement-relevant mirrorings, that is, different scale types denote different types of numerical-empirical mirrorings that are relevant for measurement. RTM, by contrast, provides the conditions under which representations on a certain scale type is possible. That is, with its representation and uniqueness theorems RTM shows what kinds of empirical systems can be represented with a structure-informative numerical system, that continues to be informative of relevant structures under specific transformations on the numerical system. In short, RTM is the formal, definitional grounding of scale type assumptions.

To illustrate RTM's foundational role in characterizing scales, consider the transitivity condition, which RTM sets as a requirement for a structure to be represented on an ordinal scale. As the biconditional formulation of the axiom indicates (see section 3.3.1), it is simply not possible to assign order-informative numbers to relations that violate the transitivity axiom. For example, consider a situation in which, for whatever way we measure and conceive of preference, Maya has strict preferences for Sacher cake over Pavlova, Pavlova over Black Forest and Black Forest over Sacher. It is evident that every assignment of numerals

[43] As evidence of the usefulness of the formal interpretation, Heilmann (2015) argues that the theorems of RTM can be used to "backward engineer" the foundations of numbers that are already in use in scientific or policy contexts. The idea is that, if we are unsure about the meaning of some numerical results we have obtained (say, scores on a test or model outputs), theorems of RTM help us investigate what kinds of objects and relations would have to exist in order for the numerical results to have an interpretation in terms of those objects and relations.

to cakes that attempts to capture that system of (strict!) order relations will fail, that is, at least one of the preference relations is not mirrored by the ordering of the assigned numerals. Our colloquial interpretation of ordinal scales as informative of order would not get off the ground without transitivity. This is how RTM provides the foundations of scale type assumptions.

Similarly, claims about interval or ratio measurement imply that a set of axioms holds, where those axioms can be proven to be jointly sufficient for constructing the relevant homomorphism. Notice though that for interval and ratio scale measurement, various axiomatizations are possible. In other words, there can be meaningful interval (ratio) level measurement of two attributes, even though the empirical relational systems they correspond to are different in some respects.

We can summarize the connection between RTM and ReMi in the following three observations:

1. Representation Minimalism is defined in terms of mirrorings.
2. Scale types denote mirrorings of different kinds of structures.
3. RTM provides formal foundations of scales by characterizing relational structures in the extension of different scales.

From these conceptual links between relational structures, scales and mirrorings, we see that RTM provides formal foundations of Representation Minimalism (ReMi). The relevance of this ReMi-RTM connection is that axiomatization from RTM can be used to evaluate the fulfilment of the requirements of Representation Minimalism. For example, if we are looking for evidence regarding representation of order relations, RTM gives us the transitivity and completeness axioms as empirical criteria. If there is empirical evidence for the fulfilment of transitivity and completeness, then there is evidence for the fulfilment of requirements of Representation Minimalism (in so far as the *specified relations* mentioned in ReMi are indeed order relations). We will see such evaluation in chapter 4, which is why the connection between ReMi and RTM is important to be aware of.

Having now carved out the relationship between RTM and Representation Minimalism, we can move on to other concerns. RTM has evoked at least two kinds of

criticisms. On the one hand, there is the worry that RTM fails to make sense of actual, successful measurement practices (e.g. Reiss 2008; Mari et al. 2017). It seems that scientists mention axioms relatively infrequently when they validate their measures, regardless of whether they want to measure temperature, time or narcissism. Having observed this about scientific practice, some philosophers argue that despite its formal elegance, RTM is useless to the actual makers and users of measures (e.g. Reiss 2008, Mari et al. 2017). On the other hand, there is debate about the exact nature of the concept "empirical relation", which is evidently central to RTM. Are empirical relations real, measurement-independent patterns in the world that are merely revealed by measurement instruments? Or is the act of measuring in some sense constitutive of the relevant empirical relations? These worries can be easily rewritten as challenges to Representation Minimalism – no wonder, given that RTM and Representation Minimalism are so closely connected. The rest of this chapter will outline how Representation Minimalism deals with these two worries.

### 3.4 Models, minimalism and validation practice

RTM has been criticized for its failure to make sense of actual, successful measurement practices (e.g. Reiss 2008; Mari et al 2017). A similar objection could be carved against ReMi: when researchers successfully validate a measure, in particular their quantitative scale type assumptions, they make no reference to mirrorings. Hence characterizing measurement-relevant representation in terms of mirrorings must be wrong-headed. In this section I want to show why claiming that ReMi is a necessary condition of measurement is compatible with the observation that successful measure validation practices indeed make no mention of mirrorings or mappings, let alone axioms.

How does one confirm scale type assumptions in practice? We can begin to look for answers in historical accounts of successful measure validation. Those histories rarely feature axioms. Rather, they involve theorizing and empirically testing law-like associations between attributes. A law-like association here means, roughly, a mathematically expressible, stable relation between the target attribute (A) and other attributes of interest (I): A = f(I1, I2, I3, … In). In psychology and social sciences, such a mathematical expression of a stable relationship between target and other (indicator) attributes is called a *measurement model* (see e.g. Embretson and Reise 2000). I will use this conception of a measurement model to

explicate two types of approaches to the confirmation of scale type assumptions: the direct axiom-based approach (DAB) and the indirect measurement model-based approach (IMM). I illustrate the two ways of confirming scale type assumptions with reference to temperature measurement.[44] Finally, I connect DAB and IMM to Representation Minimalism, showing that both approaches to scale type confirmation are compatible with Representation Minimalism. This way, I show that even though mirrorings are typically not *mentioned* in measurement practice, the practice of scale type confirmation still implicitly aims for and achieves the kind of representation Representation Minimalism requires.

When the developers of temperature measurement had achieved ordinal measurement on so-called thermoscopes, their ambitions to improve towards quantitative measurement did not direct them to axiomatic measurement theory. Rather, the focus was on the mathematical form of the law that governs the association between temperature and other relevant attributes, most importantly the association between temperature and the volume of an indicator substance, such as mercury, that fills the tube of a thermometer-to-be (under specified auxiliary conditions, e.g. pressure) (Chang 2004, ch. 2). The hypothesized model, a linear relationship between temperature and the volume of the indicator substance, achieved support gradually and abductively through a combination of theorizing and experimenting. The theorizing and experimentation eventually justified the inference that temperature could be measured on an interval scale with an instrument that is filled with the relevant indicator substance (when the instrument's cross-sectional area is constant).

Why do (approximate) confirmations of such measurement models afford inferences to measurability on a quantitative scale? Superficially, a confirmation of the linear relationship between two numerical systems (sets of numbers), one called "volume" and one called "temperature", immediately ensures that the possible numerical assignments to levels of temperature are all linearly related to each other – it could not be otherwise if the linearity of the relation between numerical systems associated with temperature and volume is confirmed. Because relations between permissible numerical systems are one crucial part of our definition of scales, we are half way through explaining how confirmation of a measurement model allows confirmation of scale type assumptions. But to comply with our

[44] The distinction between ways of confirming scale type assumptions is mine, but I rely heavily on Chang's (2004) account of the historical development of temperature measurement to explicate the distinction.

full characterization of scale type assumptions, we need something more substantial than the relation between two numerical systems: we need mappings from the permissible numerical systems to empirical relations. What does the confirmation of the measurement model have to be like in order for it to allow inferences of numerical-empirical mappings? In the case of temperature, what allows us to say that differences between numbers assigned map on to differences between entities in terms of temperature?

Notice first that in the case of temperature, the indicator attribute "volume" is known to be quantitative. Why? That volume is quantitative may seem trivial, and in fact people have treated volume as a quantitative attribute for millennia, because it is so intuitive to do so. But I will probe the foundations of those intuitions a little further here, since understanding the quantitative nature of volume will help me explain why confirmation of (some) measurement models affords inferences to scale types.

Loosely speaking, we know that the attribute volume is quantitative because we can readily observe that its subsystems are additive. That is to say, if we divide an entity into parts, the volume of the whole entity is always the sum of the volumes of the parts of the entity. (Such attributes are sometimes called "extensive" as opposed to "intensive" in the measurement literature.) Knowing this, we can mentally map the arithmetical operation of addition onto additive empirical operations involving the attribute volume. And recall that addition is a meaningful operation only for quantitative scales.

Such a chain of inference is, of course, a coarse-grained justification for the possibility of quantification. But it resonates well with mathematical measurement theories and exemplifies the core of more rigorous empirical approaches that proceed directly from RTM-style axioms. In the coarse-grained justification, one compares entities in terms of direct observations of the target attribute (e.g. how the volume of parts relates to the volume of the whole) and maps the observed relations to relations and operations on numbers (e.g. how two numbers relate to their sum) via mental association. A parallel but more rigorous empirical justification of scale type assumptions starts from measurement theory, which, as we have seen, maps uninterpreted relational systems to numerical systems via mathematical proofs. An empirical researcher can then capitalize on those proofs in her inference to scale type, if she manages to operationalize the axiomatic constraints in terms of the attribute of

interest, and if her observations confirm that the constraints are fulfilled, at least approximately.

With this in mind, we can come back to inferences from measurement models to scales. To understand that inference, I think it is helpful to recast the confirmation of the linear relation between temperature and volume (of a thermometric substance) in the following terms (recall, we are thinking about situations where the aim is to go beyond the superficial numerical-system-to-numerical-system association). What is happening, in effect, is a mapping of the known equalities and inequalities of differences in volume with the (previously) unknown equalities and inequalities in differences in temperature. In other words, we can think of the gradual confirmation of a linear law-like relation between temperature and volume (of a given thermometric substance) as the search for the conditions under which equal differences in volume map onto equal differences in temperature, and unequal differences in volume map onto unequal differences in temperature. This may sound like an unnecessarily complex way of putting the simple idea of confirming a linear law-like relation. But put this way it is easy to understand why the confirmation of the measurement model allows inferences to scale type. In the case of confirming the quantitative nature of volume, the procedure involved mapping arithmetic operations directly on empirical relations. Here, by contrast, the procedure starts from the mapping of empirical relations pertaining to volume to empirical relations pertaining to temperature, and only then proceeds to map a numerical system onto the empirical relations thus discovered. [45]

The general point of this admittedly long-winded explanation is to motivate a distinction between two methods of establishing scale types in practice: the direct axiom-based approach (DAB) and the indirect measurement model-based approach (IMM). DAB proceeds from the mathematical measurement theory to empirical reality by taking the axioms (or something like them) at face value and finding direct (that is, simple and trivially justified) observational means of checking whether the axioms hold when entities are

[45] In this example, the mapping capitalizes on the *known* quantitative properties of volume – the confirmation of the linear relation amounts, in a sense, to an extrapolation from the known quantitative properties of volume to those of temperature. However, it is *not* a general feature of model-based confirmation of scale types that one attribute must be known to be quantitative prior to the confirmation of the relevant measurement model. As we will see in chapter 4, the confirmation of a measurement model may allow inferences to quantitative scale type even when none of the constituent variables is known to be quantitative *prior* to confirmation of the model. See also subsequent footnote.

compared in terms of the target attribute. The establishment of the quantitative nature of volume is an example. In IMM, by contrast, one specifies a measurement model of the target attribute and its relations to other attributes and infers the scale type of the target attribute via confirmation of the relationship postulated in the model. The establishment of the interval measurement of temperature is an example. The division between DAB and IMM has grey areas,[46] because of the blurriness of the dividing line between direct, simple observation on the one hand and indirect, inferential means of hypothesis confirmation on the other. There is no reason to let such greyness and blurriness alarm us.

It is not a novel idea that measure validation can involve something else than direct observations of the fulfilment of axioms. Nor is it new to claim that measures can be validated via modelling. For example, Tal (2012, 2016) has defined and defended what he calls the model-based approach to measurement with respect to time measurement, while McClimans, Browne and Cano (2017) apply Tal's framework to argue that model-based considerations are central to measure validation in psychology. Much earlier, in 1934, psychologist Junius Flagg Brown wrote an article in *Erkenntnis*, where he analysed the validation of measures of e.g. weight, electrical potential and temperature in terms of the confirmation of (what in present terminology could be called) measurement models (because they express law-like tendencies) (Brown 1934). If it has all been said some 80 years ago, what has been the point of this exercise?

My aim here is to use the DAB-IMM distinction to show that seemingly different practical processes of scale type confirmation share the same conceptual framework of measurement-relevant representation. I have explicated DAB and IMM in terms of volume and temperature to show that in both cases the underlying justification for the confirmation of the relevant scale type assumption has to do with the mirroring of empirical and numerical relational systems. In the case of volume, the establishment of the mirroring is direct, consisting of the mental association of certain mathematical operations with observed

[46] From the examples I have used, one might get the impression that DAB is more fundamental than IMM, in that a DAB confirmation of the scale type of one attribute (e.g. volume) is needed to make the IMM confirmation of the scale type of another attribute (e.g. temperature). This is not a general feature of scale type confirmation, however. We will see in chapter 4 that there are IMM confirmations that do not hinge on a prior DAB confirmation. I think this detail is not essential for the IMM/DAB distinction to do work for us, which is why I shall omit extensive discussion of the point here. The point of the IMM/DAB distinction is that seemingly different kinds of confirmation activities share in the same framework of representation.

relations in terms of volume. In the case of temperature, the establishment of the mirroring is indirect, proceeding from the empirical-to-empirical mirroring of differences in the volume of a thermometric substance to differences in temperature, and from there to the numerical-to-empirical mirroring pertaining to the interval scale assumption regarding temperature. The mirroring-based notion of representation grounds both cases, even though one is an example of DAB and the other an example of IMM. In this sense DAM and IMM share the same characterization of measurement-relevant representation, even though on the surface they may seem like very different kinds of activities.

The upshot is that Representation Minimalism is compatible with common, model-based and non-model-based approaches to confirming scale type assumptions. This is good, because we know that modelling permeates psychometrics (see section 2.4). Representation Minimalism should be able to handle the modelling side of psychometrics, when we get to evaluating psychometric representation in chapter 4.

### 3.5 What are empirical relations?

A discussion on measurement-relevant representation would not be complete without a slightly more thorough look at the kinds of empirical relations numerical relational structures can legitimately mirror. Representation Minimalism is relatively non-committal in this regard: the empirical relations have to be expressed in terms of an attribute that is of interest to the measure user (e.g. observing relative lengths of rods won't do if the aim is the measurement of well-being) and the relations have to be recognizably similar to each other (e.g. representing lengths, weights and their combinations with one numerical structure won't do). This way of delineating relevant relations is rare – in fact, I have not seen it laid out in this way anywhere else. In this section I will introduce some common readings of permissible empirical relations and relate them to my reading.

Consider, as an example, the Apgar score that quickly summarizes aspects of the health of an infant. At one minute and five minutes after delivery, the midwife or the doctor assesses five easily identifiable characteristics of the baby – heart rate, respiratory effort, muscle tone, reflex irritability, and colour – assigning a value of 0 to 2 to each characteristic and summing up the scores. Total scores of 7-10 denote good condition, while scores under 7 are thought to indicate that the baby might need urgent medical attention. This prediction is

based on prior observations of the frequency with which infants that fall within a specific range of scores suffer from some serious health problems (such as brain damage). For example, studies show that the incidence of neonatal death is much higher among infants that are assigned scores 0 to 3 as compared to infants who are assigned scores 7 to 10 (e.g. Casey, McIntire, and Leveno 2001).

Does the numerical assignment Apgar score yields fulfil the requirement of measurement-relevant representation? The answer depends on what you take permissible empirical relations to be in the measurement context. In what follows I will go over some alternatives as to what Apgar score might be taken to represent, and problems with each alternative. I am *not* arguing that this or that interpretation is the correct way to read the Apgar score – I use the Apgar score merely to illustrate what kinds of relations numerical assignments might be taken to represent. I then show that my notion of representation, Representation Minimalism, is neutral with respect to these alternatives.

One might argue that the Apgar score represents empirical relations in terms of the attribute "what numbers get assigned when following the rules of the Apgar scoring system". For example, if baby Mia is assigned a "6" and baby Aija is assigned an "8", this represents the empirical relation that Mia received a lower score than Aija when they are assessed in terms of the rules of the Apgar scoring system. We might call this the *operationalist reading of empirical relations:* the numerical relational system represents the test procedure – the operation! – that yielded those numbers. One of the main problems with this approach is that measure-users are typically not interested in the procedure in and of itself, but only as far as it indicates some underlying attribute of interest to the measure-user.

To contrast the operationalist reading, one might insist on *realism about empirical relations*. On this approach, it is not enough for the numerical structure to represent the procedure that produced the numbers. Instead, the numerical structure should represent empirical relations that exist independent of the testing procedure. For example, according to the realist, the Apgar score represents (or should represent) relations between infants in terms of health, illness or well-being, where health, illness and well-being are something that exist independent of the testing procedure. The realist would typically insist that what needs to be represented is relations that bring about or cause the testing procedure to yield the numbers it yields. For example, the test-independent health status of Mia and Aija is what

causes the test procedure to assign an "8" to Aija and a "6" to Mia. The main problem with the realist reading is that it is difficult to determine whether the Apgar score tracks a plausible conception of health, illness or well-being.

We have seen the realist and the operationalist interpretation of measurement-related empirical relations. Are there others? In between realism and operationalism, one might insist, is another position: the relations that the numerical structure represents are the real, observed properties, such as skin colour and muscle tone, that the doctor or midwife reports with the Apgar score. Call this the *phenomenological reading* of empirical relations:[47] the numerical structure represents relations in terms of immediate observations, which get reported via the Apgar scoring system.[48] The main reason I will not consider the phenomenological reading here is that it does not really count as representation, that is, relations in the numerical structure cannot be interpreted or read in terms of relations in the empirical system. For example, the numbers that are assigned to Mia and Aija are not readily interpretable in terms of immediate observations, even though those numbers where assigned based on observations. The reason is that the difference in Aija's and Mia's respective scores could be due to a difference in observed muscle tone, or skin colour, or a combination of these, or a great many other combinations of observed properties. Similarly, if Mia and Aija received the same score "6", that would not be interpretable as representing the fact that Mia and Aija have a similar status in terms of observed properties. If the numerical assignment does not represent observed properties of the children, but is merely a product of considering those observations, it is not apt to consider this an approach to representation.

Another suggestion might be that alongside the operationalist and the realist interpretations, we should consider a predictive reading of empirical relations. For example, the Apgar score allows doctors and midwives to predict which infants are likely in need of

[47] The difference between the operationalist and the phenomenological reading is subtle. The operationalist says that the Apgar score represents relations between infants, when infants are compared in terms of the rules of the Apgar scoring system. The phenomonologist says that the test score represents relations between infants in terms of skin color, muscle tone and so on. On the phenomenological reading, we should be able to say how two infants differ in observed muscle tone when we see that they have been assigned scores 8 and 6, respectively. But clearly we are not justified in saying that. In the operationalist approach, we are merely saying that two infants differ in terms of their Apgar scores, which is obviously true.

[48] Borsboom's constructivist reading of RTM (2005) could be thought of as similar to what I call the phenomenological reading here. As is evident from discussions above, I deny that RTM is committed to this reading of empirical relations.

urgent medical attention. The prediction is done based on experience and empirical data on the incidence of neonatal death, brain damage and other conditions in cohorts of babies that have been assigned a specific score. For example, the incidence of neonatal death is much higher in the 0-3 range than it is in the 7-10 range. On these grounds, one might propose a *predictive reading* of empirical relations: what the numerical structure represents is (probabilistic) relations between infants in terms of their likelihood of suffering brain damage or another serious medical problem. I think this reading can have useful functions but should not be considered an alternative to the realist and the operationalist readings. This is because the predictive reading always piggybacks on either the operationalist or the realist reading, depending on what kinds of empirical relations are being predicted. In the Apgar score case, what is predicted is a test-independent attribute such as brain damage or death. In other cases, operationally characterized attributes might be predicted, for example, if one uses Facebook posts to predict scores on a depression test (and that depression test is interpreted as representing test-dependent, i.e. operational relations). Under the predictive reading, then, a numerical structure is a representation of the likelihood of the occurrence of given realistically or operationally characterized empirical relations. The adequacy of a particular predictive reading therefore largely depends on the adequacy of the relevant realistic or operational reading, which is why I shall not consider predictive reading of empirical relations as an approach of its own.[49]

The operationalists and the realists frequently argue about the relative supremacy of each position (on the opposition in psychometrics, see e.g. Michell 2008; Lovett and Hood 2011; Maul, Torres Irribarra, and Wilson 2016). The debate about the appropriate interpretation of RTM (see section 3.3.2) can be recast as a debate about whether RTM is a realist or an operationalist approach. On the one hand, the fact that Krantz et al. (ch. 1 in 1971) used length as the prime example in their explication of RTM suggests the operational reading (or phenomenological reading),[50] on the other hand the three volumes of *Foundations*

[49] Here I dealt with the idea that a numerical structure may be taken to represent predicted operationalist or realist relations. There is, of course, also the case where a numerical assignment represents some operationalist or realist relations and is simultaneously predictive of other operationalist or realist relations. I mention this distinction here just to clarify the various alternatives one might consider.

[50] The founder of operationalism (or at least its most famous champion) Percy Bridgman also used length as the showcase of operational analysis (Chang 2009). The fact that length is a key example for both the operationalists and the RTM proponents may invite the interpretation that the two are the same or very similar approaches.

*of measurement* are littered with critical remarks about operationalism (especially chapter 1 of Volume 1 and chapter 22 of Volume 3).

The beauty of Representation Minimalism is that it does not hinge on a commitment to any of these positions. This is a virtue, because picking sides at this point would considerably narrow the audience of my forthcoming arguments. I have resolved to take both the operationalist approach and the realist approach on board by characterizing measurement-relevant representation in terms of an attribute that is of interest to the measure-user. If the measure-user is a realist, the representational capacities of a measure will impress her only if the measure yields a numerical structure that represents realist empirical relations. If the measure-user is an operationalist, the representational capacities of a measure will impress her only if the numerical structure represents operational empirical relations. It turns out that, even with such weak commitments, we can get a long way in diagnosing psychometric measurement.

### 3.6 Representation in measurement

The main contribution of this chapter has been to define a notion of measurement-relevant representation. The definition I provided is:

**Representation Minimalism (ReMi)**. In measurement, a numerical representation is appropriate when specified relations in the representing numerical system mirror empirical relations between entities, when entities are considered in terms of the target attribute.

The rest of this chapter defended this definition. I firstly showed how ReMi relates to RTM, arguing that RTM provides the formal foundations of ReMi. Second, I discussed the relations between ReMi and the practice of validating claims about representation on a certain scale. I argued that ReMi is perfectly compatible with plausible accounts of how scale types get confirmed. Third, and finally, I showed that ReMi is compatible with, but more permissive than, other common notions of the kinds of empirical relations measurement aims to represent. Having defended ReMi, I shall now dicuss how successful psychometricians are at reaching a measurement-relevant representation.

# 4. Psychometric Representation

## 4.1 Old and new representational challenges

Suppose you are running a randomized controlled trial with the intention of discovering whether drug D alleviates depression. You use the state-of-the-art psychometric measure, which includes questions about suicidal ideation, depressive mood, weight loss and so on to gauge the effect of the drug on depression. To that end, you count average total scores (where total score is the result of the summation of scores on individual items) before and after the trial and compare the averages of the control group and that of the treatment group. It turns out the average total score change is what you consider small. Assume that the trial was properly randomized, the data set was large and unbiasedly handled, and so on for other support conditions, *except that the psychometric measure has not been shown to represent depression on an interval scale.* Are you justified in believing that the drug is useless for treating depression?

I believe you are not. The trouble is that without an interval level measure, a small difference between the two average score changes (one for the control group and one for the treatment group) can indicate a large difference in changes in depression (and *vice versa*). For example, it could be that people in the control group improve a little bit on a number of what might be considered peripheral or non-central aspects of depression (say, weight loss) while people in the treatment group improve significantly on core aspects such as depressive mood and suicidal ideation. Although the average score changes might be very similar, we might have good reason to nonetheless think that people in the treatment group have improved significantly with respect to depression while people in the control group have not. This is, in fact, what researchers have recently discovered about anti-depressant trials that used the extremely popular HAM-D depression measurement instrument (Hieronymus et al. 2016). Hieronymus et al. observed that the score change in the control group, even when of similar magnitude as the score change in the treatment group, tends to arise from small improvements in non-central aspects of depression, while the score change in the treatment group tends to arise from improvements with respect to depressive mood and other central aspects of depression. Whether or not Hieronymus et al. are right does not matter here. The point is that the meaning of average scores (and differences between average scores) is

opaque when score differences do not have a constant meaning in terms of the target attribute.

Validation[51] of interval scale properties, then, appears to matter in at least some common applications of psychometric instruments. Are those properties typically validated? Do we typically have evidence of interval level representation? A few commentators have argued that we do not, that is, they argue that psychometricians operate with invalid interval scale (or more broadly, quantitative) assumptions. Firstly, the proponents of RTM, such as Patrick Suppes, argue that psychometric instruments are not adequate measures, because psychometricians' scale type assumptions (including interval scale assumptions) are not validated in terms of representation and uniqueness theorems (e.g. Suppes and Zinnes 1962, Krantz et al 1971, ch. 1). Joel Michell pursues a similar argument in several articles (e.g. Michell 2008), arguing that psychometricians operate as if they have evidence that their target attributes are quantities (which implies they are representable on an interval scale), although that assumption has not been corroborated by evidence. I shall call these *representational challenges:*[52] they challenge the supposition that psychometric instruments yield quantitative representations of their target attributes.

How have representational challenges fared? Have they changed anything? No. For example, most psychometricians are not aware of RTM and its proponents' critique of psychometrics. Several kinds of reasons have been offered to account for the fact that these representational critiques have had little bite. For instance, commentators such as Norman Cliff note that the formal apparatus of RTM is foreign and impenetrable to most psychometricians (Cliff 1992). Others note that there is almost no "bridging work" that makes the formal results more accessible (see also Luce and Narens 1987; Luce 1997a). In other words, the claim is that the representational challenges have not been taken seriously because they have not been understood.[53]

---

[51] Note that I use the term "validate" broadly in this chapter, not adhering any specific reading we encountered in chapter 2. To validate means to establish or confirm the truth of some claim, in this case a claim about the property of a measurement instrument.

[52] Although Michell's critique might be better thought of as a realist critique, as discussed below.

[53] This claim is plausible, especially since RTM is often *not* understood to define representational requirements for all measurement. Rather, it is treated as one among many approaches that provide a methodology for the validation of measures (Judd and McClelland 1998 to an extent; John and Benet-Martinez 2000; Angner 2011). In brief, RTM is sometimes taken as an (overly formal and impracticable) alternative to achieving the same thing psychometrics is trying to achieve (valid measurement), rather than as a theoretical framework for all of measurement.

Others argue that the representational challenges have not only been poorly understood, but that they are also mistaken criticisms. Hence psychometricians can safely ignore the challenge, whether or not they comprehend it. First, some argue that the representational critique is based on an overly restrictive view of what measurement requires. We saw this criticism pop up in many different forms in the previous chapter, where we discussed RTM. Others reject the representational challenges on different grounds: they argue that the representational critique applies only to a very narrow conception of psychometrics. For example, Borsboom and Mellenbergh (2004) argue that at least some of the above-described criticisms do succeed in undermining Classical Test Theory, but that for example Rasch measurement is immune to the representational challenge, because Rasch measurement does provide evidence of interval level representability.

In this chapter I present my own representational challenge: The New Representational Challenge. It builds on earlier representational critique(s) but is also informed by the sceptical responses representational critiques have thus far received. The New Representational Challenge can be summarized as follows:

**Premise 1**. Psychometric instruments are often treated as if they yield an interval representation[54] of the target attribute.
**Premise 2**. Psychometric validation usually fails to validate an interval representation of the target attribute.
**Conclusion.** When psychometric instruments are used, the assumed representation is usually not validated.

The new challenge improves upon the previous representational challenges in several ways. Firstly, the New Representational Challenge is *not* presented formally, unlike RTM and much of Michell's work is. This makes it more accessible than earlier critiques.[55] Second, where possible, the new challenge is expressed in "psychometricians' own terms",

[54] "Interval representation" is short for "representation of the target attribute on interval scale".
[55] This is not to say that psychometricians are not good with formal arguments. I am just picking up other authors' observation that the formal apparatus representational challenges are typically presented with is not familiar to psychometricians. More generally, formal presentation has the problem that it leaves room for various empirical interpretations, which may lead to misinterpretation. A conceptual explanation of representational problems is both new (in this context) and helpful (in my view).

that is, taking on board the representational assumptions psychometricians themselves make. This move is important to disarm the critique that representationalists apply an overly restrictive criterion of measurement (to a field that does not recognize the usefulness of such a restrictive view). Third, I will apply the new challenge to both classical and modern psychometric approaches, to avoid the objection that representationalists bash an overly narrow version of psychometrics. Fourth, the new representational challenge is metaphysically less demanding than some other representational challenges. In particular, it is less demanding than the realism proposed by Michell (e.g. 2005), which implies, among other things, that the real numbers are "spatiotemporally located relations". I have deliberately adopted minimal constraints on what counts as an acceptable target attribute for the purposes of measurement (chapter 3, esp. section 3.2) – in particular, I have not committed to realism about attributes.

Finally, I depart from many other critics of psychometrics when it comes to the degree of hopelessness I think the representational critique should evoke. Many critics of psychometrics regard their criticisms as more or less death sentences to psychometrics as we know it. Michell (2008) famously calls psychometrics pathological science and prescribes a whole new philosophy of measurement as a remedy. Similarly, RTM proponents' disappointment with psychometrics leads them to formulate a formal measurement theory, which resonates so poorly with psychometricians that the theory has been almost entirely side-stepped (by psychometricians). My conclusion, however, is not that psychometrics is in a hopeless state. Chapters 5 and 6 will explain relatively simple interpretational moves that allow us to appreciate the usefulness and meaningfulness of psychometric instruments, the conclusions of this chapter notwithstanding.

The structure of this chapter is the following. Sections 4.2 and 4.3 establish Premise 1. Section 4.4 establishes premise 2. Section 4.5 deals with the implications of the argument. Section 4.6 concludes.

### 4.2 Target attributes are non-operationally defined

As per chapter 3, target attribute is the attribute the measurer is interested in measuring. How might a single general argument take on board the myriad of attributes psychometric instruments are used to measure? I need one weak, and (in my view) plausible,

assumption about the kinds of attributes psychometricians are interested in: non-operationally defined target attributes. More precisely, then, the argument of this chapter has the form:

**P1.** Psychometric instruments are often treated as if they yield an interval representation of a non-operationally defined target attribute.
**P2.** Psychometric validation usually fails to validate an interval representation of a non-operationally defined target attribute
**C.** In psychometrics, the assumed representation is usually not validated.

An operationally defined target attribute means (here) an attribute that is defined solely in terms of the operation for measuring the attribute. An operational definition of an attribute is of the form:

"A is whatever M measures"

For example:

"Intelligence is whatever the Stanford-Binet test measures."
"Well-being is whatever the Satisfaction with Life Scale measures."
"Depression is whatever the Hamilton Depression Scale measures."

Many people have been suspicious of operationally-defined attributes. The intuition behind such suspicions tends to be that the operationalist approach defines away central questions of measure-making, and that defining problems away is an unsatisfactory way to resolve problems. For instance, consider the following questions that I take to be central to measurement:

"Does measure M really measure intelligence?"
"Is measure M a valid measure of intelligence?"

The usual objection to operationalism is that when attributes are operationally defined, these questions become either meaningless or their meaning is significantly distorted from their usual meaning.[56] The first question is usually considered non-trivial, in the sense that it is a genuine possibility that measure M does *not* measure intelligence. But on the operationalist approach (so the critics say) the answer to the question is always, and trivially, "yes", because intelligence is defined by measure M. The second question is usually taken to connote non-trivial subquestions such as "Are the results accurate?", "Is the measure tracking intelligence rather than some related attribute like verbal ability?" and "Is the measure based on a defensible conception of intelligence?". But it appears that in the operationalist approach, all these subquestions must be answered yes, whatever the measure. More generally, it appears that in the operationalist approach, a measure is valid by construction – unless validity means something else than what it is usually taken to mean. Critics of operationalism regard it as illegitimate to define away the need for validation.

So, there are some general considerations that favour non-operationally defined attributes. Do psychometricians share these anti-operationalist intuitions? The issue is somewhat contentious because psychometricians are often blamed for blind operationalism (e.g. Green 1992; Michell 2008). Many psychometric textbooks seem to indeed recommend, at least nominally, some sort of operationalism (see Green (1992) for a review of this phenomenon). Nonetheless, I think psychometricians tend to want to measure non-operationally defined target attributes. This desire seeps into psychometricians' practices and arguments, whatever they might explicitly say if prompted to analyse the nature of their target concepts (see Leahey (1980) for a similar analysis). The following is a list of practices that characterize much of psychometrics and that testify to the desire and attempt to measure non-operationally defined attributes.

**Exhibit A: Controlling for bias.** Psychometricians have a keen interest in detecting and controlling for biased rating (Saal, Downey, and Lahey 1980). In the context of rating instruments, bias denotes, roughly, people's tendency to give ratings that do not reflect their standing on the target attribute. In particular, bias concerns people's tendency to give

[56] These claims about "usual meanings" of questions about measurement are based on my own observations of how people talk about measurement in and outside of academia.

scores based on their "preference" for certain ratings independent of their standing on the target attribute. Consider, as examples, the following common biases:[57]

**Severity:** subjects give low scores independent of their standing on the target attribute.
**Leniency:** subjects give high scores independent of their standing on the target attribute.
**Everybody is average – bias**: subjects always choose ratings midway the rating scale.
**Halo bias:** subjects' scoring on all items is influenced by a single item they consider salient – if they score low (high) on that salient item, they will give low (high) ratings on all other items as well.
**Social desirability bias**: subjects rate in a manner they expect others would view favourably.

There are a number of other biases (or rater effects), the details of which do not concern us here. The point is that psychometricians' interest in bias, and the efforts to control for it, would not make sense if psychometricians were content with operationally defined attributes. If the attributes were operationally defined (in the sense described above), the ratings could not fail to reflect the target attribute, so the concept of biased rating would be an oxymoron. But if psychometricians thought bias was an oxymoron, they would not be as keen to detect it and control for it as they actually are.

**Exhibit B: Construct validity.** The most celebrated psychometric notion of validity, construct validity, is typically associated with the measurement of non-operationally defined attributes. This is without a doubt the implication when Loevinger (1957, 642) says that:

> *Construct connotes construction and artifice; yet what is at issue (in construct validation) is validity with respect to exactly what the psychologist does not construct: the validity of the test as a measure of traits which exist prior to and independently of the psychologist's act of measuring.*

[57] The literature is not entirely consistent regarding definitions of different biases (see Saal, Downey, and Lahey 1980).

Similarly, Cronbach and Meehl (1955) could not be clearer when they write that:

> *Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not "operationally defined".*

Whatever one thinks of the fruitfulness of the concept of construct validity, it is evidently the most talked-about and valued notion of validity in much of the psychometric literature. The fact that classic statements of construct validity define it in terms of non-operational target attributes is evidence that psychometricians value and aim for the measurement of non-operationally defined target attributes.

**Exhibit C: Comparison of measures.** It is common for psychometricians to compare different measures vis-à-vis their ability to capture the target attribute. For example, consider a review of measures of subjective well-being written by Ed Diener, the developer of the famous Satisfaction with Life Scale (SWLS). Diener writes that single-item measures "cannot hope to cover all aspects of" subjective well-being the way multi-item measures do (Diener 1984, 544). By this he means that a single rating-item cannot gauge the various aspects that constitute subjective well-being, such as different affective and cognitive dimensions. Such comparisons reveal that (Diener thinks that) each measure of subjective well-being does not define a new type of target attribute. Otherwise it would not make sense to assess different measures on their ability to track a specific, multidimensional subjective well-being. Similarly, measures of depression, such as HAM-D, Beck Depression Inventory and Montgomery Åsberg Depression Rating Scale are frequently compared to each other in terms of e.g. their "precision in estimating depression" (Carmody et al. 2006). Such comparisons can be interpreted in many ways, but one salient reading is that the goodness of a measure of depression is being evaluated in terms of its ability to capture depression non-operationally defined.

**Exhibit D: Latent variable talk.** The notion of latent attribute (or latent variable) is part and parcel of much of psychometric work, but the concept has multiple possible interpretations. Many psychometric validation practices imply, explicitly or implicitly, a non-

operational understanding of latent attributes, i.e. something existing independent of the test set-up. Borsboom, Mellenbergh and van Heerden (2003) argue, for example, that a consistent interpretation of the psychometric usage of latent variable models (e.g. IRT models) requires a realist interpretation of latent variables. By this they mean, roughly, that the way psychometric models are used implies adherence to the idea that the target attribute (i.e. the latent attribute) exists in reality, whether or not we measure it (Borsboom, Mellenbergh, and van Heerden 2003).

Adherence to such "realism" is not just implied, but also explicitly endorsed by many psychometricians. For example, in a relatively recent paper published in the *Proceedings of the National Academy of Sciences*, the authors write thus about their notion of personality as a latent variable:

> *Personality traits, like many other psychological dimensions, are latent and cannot be measured directly […]. We adopted the realistic approach, which assumes that personality traits represent real individual characteristics…* (Youyou, Kosinski, and Stillwell 2015, 1036)

Of course, definitions of realism vary within philosophy, let alone across sciences. But one plausible reading of the above quote is that the target attribute is not defined in terms of the measurement operation, but rather in terms of a test-independent reality.

In sum, there is much to suggest that psychometricians frequently target non-operationally defined attributes. Of course, this is not always the case. Sometimes a researcher's sole interest is the observed responses, not a non-operationally defined attribute that drives those observed responses. For example, consider a researcher studying people's judgment of different kinds of moral wrong-doings. She may be interested solely in, say, the observational level question: "Do people pass different judgments on different moral wrong-doings" rather than a deeper, attribute level question, say, "What do people's observed moral judgments mean in terms of the attribute moral leniency?" In other words, her focus is the judgments for their own sake, not as indicators of some underlying attribute that drives those

responses. Such cases are not instances of treating the instrument as a measure of a non-operationally defined attribute.

In chapter 6 we will look at ways to defend operationally defined target attributes. For now, we will proceed to evaluate parts of psychometrics that commit to non-operationally defined attributes.

### 4.3 Intervals are in use

Try taking the average of a ranking. That is, the average of a ranking that does not convey information about the magnitude of difference between different positions in the ranking. Say, count the average beauty of 10 Miss Universe runner-ups based on their ranking in the contest. Or the average age of children in a family based on their ordering from oldest to youngest: Henu ≻Rei ≻ Kailo. Or the average hardness of minerals based on their order on the Mohs hardness scale.

Taking averages and treating them as informative of the average incidence of an attribute is a fairly sure sign that a stronger than ordinal representation is presumed, either implicitly or explicitly. In the common classification of scales, interval scales are the weakest scales that allow meaningful statements about averages, that is, meaningful in the sense that the average of the numbers assigned can be interpreted as the average incidence or occurrence of the attribute of interest in the population of interest.

Unsurprisingly, it is common for psychometricians to study averages of ratings and interpret them in terms of the target attribute. Consider, for example, the HAM-D rating scale, which is intended to gauge levels and changes in depression. HAM-D typically consists of 17 observer-rated items, with different associated rating-scales. For example, the following item is intended to assess the subject's feelings of guilt:

FEELINGS OF GUILT

0 = Absent

1 = Self-reproach, feels he/she has let people down

2 = Ideas of guilt

3 = Present illness is a punishment; delusions of guilt

4 = Hallucinations of guilt

Ratings from different items are summed to form the total HAM-D score of an individual.

HAM-D is frequently used in drug trials to evaluate the efficacy of an anti-depressant. In particular, change in the average total HAM-D score is frequently taken as one of the main indicators of the efficacy of the drug (e.g. Khan, Warner, and Brown 2000; Kirsch et al. 2002, 2008). The researchers compare changes in the average HAM-D scores for treatment and control groups, and if the magnitude of change in the treatment group's average total score is significantly higher than that in the control group, that is taken as evidence for the efficacy of the drug as a treatment of depression.[58] In other words, changes in average scores are interpreted in terms of average changes in the attribute of interest, that is, depression. Using averages in this way suggests commitment to an interval level interpretation of HAM-D scores, because averages of ordinal scores would not have any interpretation in terms of the target attribute.[59]

It is, of course, not just averages that reveal the assumption of interval representation. Other indicators include, for example, the addition of scores across items to form the total score and then interpreting the total score in terms of the target attribute, and the comparison of magnitudes of changes on item and total scores as indices of changes on the target attribute.[60] These practices are common, as one might suspect. In the HAM-D case, the total score of a subject is formed by simple addition of scores across items that track things like insomnia, feelings of guilt, suicidal ideation and weight loss. Hence two people can have the same total HAM-D score, but that score is constituted of very different item scores (say, one gives high scores on guilt and insomnia and low scores on suicidal ideation and weight loss, and *vice versa* for the other subject). Treating the subjects' total scores as their standing

[58] There are other ways of detecting effectiveness of a drug, such as defining effectiveness in terms of 50% score reduction from start to finish of the trial.

[59] One might perhaps argue that HAM-D users define depression operationally, and therefore the interval scale assumption is not as controversial as I am suggesting here. This strikes me as an implausible objection. If HAM-D was thought to define depression, debates about the non-centrality of certain items, and more general concerns about the validity of the instrument would not make much sense. One would also expect a very different kind of communication about the significance of results of HAM-D if HAM-D was taken to define depression. In any case, my argument does not hinge on this or that interpretation of HAM-D, because I use HAM-D merely to illustrate how one can detect interval scale assumptions.

[60] For more technical discussions of (statistical) techniques that presume interval representation, see Maxwell and Delaney (1985); Embretson and Reise (2000 ch. 6).

on the attribute depression suggests that all one-point score differences across items are intended to have the same meaning in terms of depression.

In brief, then, prevalent score treatment suggests that:

**P1.** Psychometric instruments are often treated as if they yield an interval representation of the target attribute.

Many other authors have noted that this assumption is common in psychometric practice (Coombs 1950; Cliff 1989; Blanton and Jaccard 2006; Borsboom and Scholten 2008; Kristoffersen 2010; Furr 2011, chap. 2). Some of the textbooks discussed in chapter 2 even testify explicitly that interval scales are commonly assumed in psychometrics (e.g. Lord and Novick 1968, sec. 1.6). The interested reader can explore manifestations of this phenomenon via the above-listed literature.

For the sake of completeness, it must be mentioned that the usefulness of results from psychometric measures does not always require assuming interval level interpretation of scores. Indeed, many psychometricians who reflect on the matter (which is arguably not many) feel uneasy about committing to an interval level interpretation of ratings. They have therefore sought ways to utilize the scores without this commitment. For example, so-called standardization is thought to bestow meaning on numerical data without committing the user to an interval interpretation of scores. These techniques are useful, but they do not challenge the claim that, when psychometric measures are used to represent their non-operationally defined target attributes, the assumed representation is often an interval level representation.

To illustrate why this is so, let's zoom in on a popular standardized effect size measure, Cohen's d. d is simply the difference between the mean score (or score change) for the treatment group and that of the control group divided by the standard deviation of scores (or score change) across groups.[61] Formally, Cohen's d is calculated as follows:

[61] The relevant standard deviation is sometimes called "pooled standard deviation", which denotes the weighted average of standard deviations for the two groups.

$$\frac{M_X - M_Y}{\sigma_{X,Y}}$$

The main motivation for studying d is the need to interpret differences between groups in terms of a shared yardstick or "unit". The shared yardstick is the standard deviation: dividing the difference in mean scores by the standard deviation allows the researcher to contextualize the mean score difference in terms of the spread of the data. This contextualization is taken to be more meaningful and easier to interpret than the simple difference in mean scores.

d is useful for gaining a grasp of the way two score changes relate to each other. But importantly, d is not a measure of the target attribute. This is because the "shared yardstick" in d is *not* the target attribute but the spread of the data. In other words, the magnitude of difference d expresses is not the magnitude of difference on the attribute of interest, but rather the magnitude of difference in terms of the spread of the data. Consequently, equal difference on the attribute of interest can give rise to different ds in two studies, and *vice versa*, the same d can arise in two studies where the difference on the attribute of interest is unequal.

Consider, for example, a situation where the target attribute is weight loss in two studies of diet pills. The difference in average weight loss between treatment and control group can be exactly the same in the two studies, but d is different, because the distribution of the weight loss is different in the two studies. Conversely, d may be the same in the two studies, even though the difference in weight loss between treatment and control groups is different in the two studies. Clearly, d is not readily interpretable in terms of the target attribute (weight loss), even when we are dealing with an attribute that has a known interpretation in its non-standardized form.[62] It is therefore not true that d represents (differences on) the target attribute without committing the researcher to an interval scale interpretation of scores, because, without additional interpretative moves, d does not represent the target attribute to begin with!

Of course, one often needs to use d to make inferences about the target attribute. When d is interpreted in terms of the target attribute, the question about the

[62] On the interpretation of effect size, see e.g. Blanton and Jaccard (2006); Lakens (2013).

justifiability of interval scale interpretation of scores rears its ugly head. Why? An interpretation of d in terms of the target attribute can only come in two forms: i) the average scores that go into the calculation of d are interpreted in terms of average levels of the attribute, in which case we are back with the problem of justifying interval interpretation of scores, or ii) differences in d are interpreted in terms of differences on the target attribute independent of raw score interpretation, in which case the question of justifying interval scales, all but avoided, is simply applied to a different numerical assignment (i.e. the assignment of d rather than the assignment of average scores). In sum, while the study of d can provide useful insight, it does not free us from justifying interval scale interpretation of results of psychometric measures.

### 4.4 Intervals are not validated

We have now established Premise 1:

**P1**. Psychometric instruments are often treated as if they yield an interval representation of a non-operationally defined target attribute.

In this section we turn to establishing Premise 2:

**P2.** Psychometric validation usually fails to validate an interval representation of their target attribute.

Before we embark on arguing for the truth of premise two, a disclaimer is in order. As I have shown in chapter 2, there is a great variety of techniques, and interpretations of the results of those techniques, that go under the title “psychometric validation”. I cannot go through all these varieties. Rather, I will discuss common versions of each technique to demonstrate my findings.

In particular, I will not talk about the interconnections between techniques, and how they can support each other. For example, I will not discuss how inter-item correlations and inter-test correlations bear on each other. The justification of this aspect of my study is that it is relatively rare for psychometricians to support a measure’s claim to validity by

appealing to interconnections between different observed measurement properties. That is, while psychometricians do run different tests of measure properties, e.g. reliability, validity, model-fit, they rarely interpret, say, reliability indices in light of validity indices, or vice versa. Rather, each aspect is scrutinized as an independently important aspect of the measure-to-be. We received some evidence for this in chapter 2, when I argued that the expansive notion of construct validity (where various statistical tests are interpreted with respect to each other and a theory) is not commonly acted upon in practice. It is even less common that results from checks of reliability, validity and model-fit would be compared to each other to support claims about scale type assumptions.[63] Hence, I will evaluate the potency of *individual techniques* to yield the interval scale representation psychometricians tend to aim for. Interconnections between measure properties will be discussed further in section 4.5.

### *4.4.1 Coefficient alpha*

Coefficient alpha is by far the most common indicator of measure reliability in the psychometric literature (Cortina 1993). As you may recall from chapter 2, coefficient alpha is defined as follows:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum \sigma_{Y_i}^2}{\sigma_X^2}\right)$$

where n is the number of items,

$\sum \sigma_{Y_i}^2$ is the sum of the variances of scores on items $Y_1$, $Y_2$, … $Y_n$ in test X, and

$\sigma_X^2$ is the variance of the total test score on test X.

To assess coefficient alpha in terms of its contribution to knowledge about quantitative representation, imagine the following test setting. We have a 5-question questionnaire with a 7-point scale, which respondents have to use to indicate their agreement

[63] There is an abundance to choose from, but here are examples of articles that do not interpret various statistical tests vis-à-vis each other and in terms of quantification: Diener and Pavot (1993) of the Satisfaction with Life Scale; Bagby et al. (2004) on the Hamilton Depression Rating Scale; Tarescavage et al. (2015) on the Minnesota Multiphasic Personality Inventory-2-Restructured Form; Rammstedt and John (2007) on the 10-item Big Five Inventory.

with each test item (let 7 stand for complete agreement and 1 for complete disagreement). Such a setting matches neatly with e.g. psychologist Ed Diener's Satisfaction with Life Scale (SWLS), where each of the five items is intended to gauge a person's satisfaction with life (Diener et al. 1985). For example, one of these items asks whether the subject agrees with the statement: "I am satisfied with my life".

Imagine also that having collected responses to the SWLS questionnaire (or another psychometric rating instrument), we arrive at the data set that I have summarized in Table 8. The data set is obviously too small for meaningful inference about properties of a test but it suffices for illustrating how coefficient alpha works. For a data set like this, alpha is over 0.9, which is very high.

| | **Q1** | **Q2** | **Q3** | **Q4** | **Q5** |
|---|---|---|---|---|---|
| **Sisu** | 7 | 6 | 7 | 6 | 7 |
| **Rei** | 5 | 4 | 4 | 5 | 4 |
| **Kailo** | 2 | 1 | 2 | 1 | 1 |
| **Jalo** | 5 | 4 | 4 | 5 | 4 |

*Table 8. Simplified example of test results that yield a high coefficient alpha.*

In simplified terms, what coefficient alpha tracks is the consistency of the pattern of responses on different items. A high coefficient alpha indicates that the patterns of responses tend to be (in certain respects) similar on different items. In the above table, for example, we can compare patterns of responses on any two items and see that for any subject the two responses tend to have the same direction vis-a-vis the item mean. Accordingly, alpha is high.[64]

[64] We can grasp this intuition in terms of the formal definition of alpha as well. The denominator in $\frac{\sum \sigma_{Y_i}^2}{\sigma_X^2}$ tends to be relatively small when individuals have an inconsistent pattern of scoring and relatively large when individuals have a consistent pattern of scoring (relative here means relative to the nominator). This is, very briefly, because individuals' total scores tend to converge with each other when individuals score high on some items and low on others (inconsistency), while the total scores tend to diverge when some individuals give consistently high scores across items and others give consistently low scores across items. Thus, with an

With the above definition in mind, let's focus on the question: does coefficient alpha vouch for representation of the target attribute? That is, does coefficient alpha help us determine whether the measure represents the attribute of interest rather than some other attribute? It does, but in a preliminary manner: information on whether responses on different items vary mutually consistently indicates how plausible it is that different items track the same or a closely related target attribute. One rationale for this is that if people tend to respond in a similar fashion to different items, they likely interpret the different questions as pertaining to the same or closely related attributes. If, for example, all the questions on Diener's SWLS prompt you to evaluate the same (or closely related) aspect of your life, you are likely to give high (low) scores on all questions if you consider yourself high (low) on that aspect. Frequently, measurers are interested in a single target attribute, or at any rate a cluster of conceptually closely related attributes – one rarely wants, for example, a combined measure of intelligence and physical strength.[65] In such typical circumstances, a measure can hardly track the target attribute if its different components track different attributes. Hence coefficient alpha does bear on the representation of the target attribute.

The contribution to capture is "preliminary" (but not unimportant), because a high coefficient alpha says nothing about which attribute is being consistently captured. The subject's ratings do not inform us about the kinds of considerations that prompted a consistent or inconsistent response pattern across items, that is, we do not know what specific attribute the subject evaluated when giving her rating. For example, questions on Diener's SWLS might prompt one to think about overall health over one's lifetime, one's economic situation, (average) intensity or frequency of positive emotions one experiences, number of life goals one has fulfilled, and so on and so worth. It is therefore appropriate to follow Fiske (1971) in thinking that coefficient alpha indicates "the extent to which the several items are measuring the same quality [or related qualities] in the individual subjects" but not what that quality is.

Coefficient alpha does not, however, provide evidence regarding interval level representation. This is because the internal considerations that determine subjects' ratings

---

inconsistent pattern of scoring, the whole expression $\frac{\sum \sigma^2_{Y_i}}{\sigma^2_X}$ will get a relatively large value, which in turn translates into low alpha.

[65] I will comment on measures that track various attributes below.

are unknown and likely differ from subject to subject in unknown ways. In particular, they are influenced by bias in ways that coefficient alpha does not (and is not mean to) reveal. For example, it could well be that Sisu, Rei and Kailo fare exactly the same in terms of life satisfaction (however construed except bearing in mind the commitment to non-operational definitions), but due to leniency, the "everybody is average"-bias and severity, respectively, they end up giving different scores on each question (as presented in table 8). Similarly, Rei might in fact be more satisfied with her life than Jalo, but since one or both of them suffer from the "everybody is average"-bias, they give exactly the same ratings. The scores thus do not represent order relations, let alone stronger relations required for quantitative representation. Calculation of coefficient alpha does not (and is not meant to) uncover and correct for such rater effects – it taps simply into the consistency of the observed scores.

Before we move on, I should mention that sometimes measurers want to capture different target attributes with the same measure, or at any rate they do not mind if different items track different attributes or dimensions. For example, the HAM-D might be taken to track different dimensions such as sleep disturbance, somatic symptoms, depressive mood and anxiety-related symptoms. In such cases researchers may not require a very high coefficient alpha, because different items can justifiably exhibit divergent response patterns, if they are intended to measure different things. For our purposes it is important to note that it is extremely difficult to make the case that the multidimensional target attribute is represented on an interval scale.[66] This is because with a measure that tracks multiple dimensions simultaneously, a difference between two instrument readings can be due to a difference in terms of any of the measured dimensions or a combination of dimensions. Therefore, differences between numbers do not have a single but multiple interpretations. That, in turn, means that equalities of differences between numbers do not signal equalities of differences in terms of what is being represented. But that is what interval scales require: (in)equalities of differences between numbers having an interpretation in terms of (in)equalities of differences in what is being numerically represented.[67] The upshot is that the

[66] That measurements should always pertain to a single attribute (i.e. a unidimensional target) is an idea that occurs in many places in psychometric literature (and plausibly elsewhere in measurement literature). Horton et al. (2013) assign this position to e.g. Thurstone.

[67] The literature on homogeneity, which we touched upon in chapter 2, is pertinent here. Many psychometrics experts make the same point I am making here, arguing that homogeneity is crucial for quantitative

justification of interval scales is particularly hard to achieve with measures that are not even intended to yield high coefficient alphas.

### *4.4.2 Construct validation*

Although construct validation comes in many guises, as we have seen in chapter 3, for the purposes of this evaluation we may focus on two common elements: i) construct validation involves checking correlations between different measures, and ii) the convergences and divergences that are discovered in such correlational evidence are interpreted in light of a theory about the target concept. The typically utilized measure is Pearson correlation coefficient, which is usually expressed as follows:

$$\rho_{X,Y} = \frac{E[(X - M_x)(Y - M_Y)]}{\sigma_X \sigma_Y}$$

where $\sigma_i$ is standard deviation on test i,

$M_i$ is mean score on test i,

X and Y are test scores.

As an example of theory-referenced correlational testing, consider the validation of the aforementioned SWLS. When discussing the construct validity of the SWLS scale, Diener and Pavot (1993) write that ”the SWLS has been found to be positively correlated with extroversion and inversely correlated with neuroticism [references omitted], thus adding to the construct validity of the scale” (Pavot and Diener 1993). On the present reading of construct validity, such patterns of correlations “add to the validity of the scale” because there are (allegedly) theoretical reasons to expect such patterns. For example, Diener and Pavot suggest that the theoretical underpinning of the expectation that extroversion and subjective well-being should be positively correlated is that “extroverted individuals have more sensitive reward systems” than those who are not extroverted, therefore making them more likely to have high subjective well-being.

representation. The interested reader can turn to e.g. Thomson (1940), Loevinger (1947) and Michell (2012) for details, for we cannot explore this literature further here.

Setting again aside, for the moment, the question of *type* of representation (ordinal, interval, ratio), let's first ask how construct validation contributes to ensuring that the measure yields representations of the attribute of interest (rather than some other attribute). Construct validation does provide means to ensure this. If we have theoretical reasons to expect e.g. the co-occurrence of extraversion and subjective well-being, and our proposed measures of extraversion and subjective well-being have a positive correlation, that is some kind of evidence for the ability of the measures to capture the target attributes. When more theoretical interconnections get confirmed, we can justifiably become more convinced that of our measure tracks the correct attribute, because it would be extremely unlikely that the observed intercorrelations occur by accident. On the other hand, when correlational evidence and theoretical expectations fail to cohere, the implications are underdetermined: the problem might lie in the theoretical expectations or the measure or in any number of methodological choices and background assumptions that go into a measure validation exercise. Now, underdetermination is no news, and while it does complicate inferences to measure validity, it does not undermine the claim that construct validation yields at least some (although not infallible) evidence for capture of the correct target attribute.

Recalling rater effects, a pessimistic evaluator might argue that high correlations between two rating instruments are likely due to the persistence of rater effects: a lenient (severe) rater will give high (low) scores on both measures, not because she is consistently evaluating related attributes, but because she is consistently lenient (severe) whatever items she faces. Hence high correlations on related measures are not evidence for capture of the intended target concept. This worry is mitigated by evidence of inverse correlations (e.g. subjective well-being is inversely correlated with neuroticism) and evidence that the proposed measure of our target concept has no significant correlation with measures of other concepts, which are not conceptually or otherwise related to the target concept.

What about scale types? The correlational evidence is likely to bear on ordinal representation, that is, appropriate divergences and convergences between measures tell us something about subjects' ability to rank their status on the attributes of interest. To see why, consider the following statements:

**A1.** Our theoretical expectations about the target concept are correct (e.g. neuroticism, extraversion and subjective well-being have the postulated relations).

**A2.** Responses on all rating instruments conform with the theoretical expectations.

**A3.** On all the rating instruments of interest (e.g. subjective well-being, extraversion, neuroticism), it is common that subject x gives a higher score than subject y, but x has less of the attribute than y.

**A4.** All rating instruments of interest (e.g. subjective well-being, extraversion, neuroticism) capture the correct target concept.

It is easy to see that, if A1 and A2 are assumed true, A3 and A4 cannot both be true. In other words, it is not possible that people are consistently *incompetent* at rating their interpersonal ranking on an attribute of interest and that the measures still successfully tracks the attribute of interest. The upshot is that, holding our theoretical assumptions fixed (i.e. A1 is true), the evidence from a successful construct validation exercise (i.e. A2 is true) simultaneously provides evidence for capture (i.e. A4 is true) and the hypothesis that people tend to use the rating instruments to competently report their ranking on the target attribute (i.e. A3 is false). In sum, evidence of capture of correct attribute goes hand in hand with evidence of competent ranking vis-à-vis the target attribute.[68]

An important caveat here is that sometimes rankings are more informatively interpreted in terms of categories. Thus, a construct validation exercise could be successful (i.e. the expected correlations emerge), but the appropriate interpretation of the instrument readings is nominal rather than ordinal. Consider, for example, the SWLS question: "If I could live my life over, I would change almost nothing". It could be the case that people typically respond to this question as if it was an either/or question: they choose a rating above 4 if they consider they would change almost nothing, and below 4 if they would change some things. In other words, what matters is the rating's relation to the midpoint and the category that midpoint-relation is taken to signify (no change/change), not the ranking the midpoint-relation is taken to signify. Importantly, a successful construct validation exercise on its own does not exclude the possibility that a categorical interpretation fits better than an ordinal

[68] Underdetermination seeps in again, because clearly one cannot assume the truth of the theoretical expectations. Underdetermination in construct validation can be studied from e.g. Cronbach and Meehl (1955).

one, because whether people interpret questions in terms rankings or categories is a black box that construct validation does not probe.

When it comes to quantitative representation, construct validation is pretty much toothless.[69] Against the backdrop of a growing literature on rater effects, it seems highly unlikely that subjects assign the same meaning to equalities and inequalities of differences between scores. That is, the change in life satisfaction that would make me choose 5 over a 4 on any of the SWLS questions is likely not the same "amount" of life satisfaction that would make me, let alone someone else, choose 2 over 1. And nothing about the standard correlational evidence that construct validation relies on can reveal the interpretations people assign to differences between scores, when they consider their status on the attribute of interest.

Even a linear relation between two measures does not establish interval level interpretation of scores in terms of the target attribute.[70] That is, a (approximately) linear relation between responses on two rating instruments can emerge whether or not the ratings track equalities and inequalities of differences in terms of the target attribute. For example, two measures might be approximately linearly related if they ask very similar questions using a similar rating-prompt. To give a crude example, consider a situation where people face two items:

I am happy, and

I am at least somewhat happy,

and are asked to rate their standing on the previously introduced 7-point scale in response to both statements. It is likely that most people will give a slightly lower rating on the first question, because it is easier to be confident about one's agreement with the cautiously worded item than it is with the big, philosophical-sounding claim "I am happy". Perhaps most will opt for just one "point" lower rating – at any rate, let us imagine that this is the obtained response-pattern, for the sake of this illustration. These two one-item measures would

[69] As argued in chapter 2, construct validation is a malleable concept that can be taken to denote many different things. One could therefore argue that construct validation *qua interpretation of correlations vis-a-vis a quantitative theory* does serve the establishment of interval scale representation. I have never seen construct validation conducted in terms of a quantitative theory with the aim of confirming scale type assumptions. That is why I will not consider this hypothetical form of construct validation here.

[70] Recall that interval scales are defined in terms of linear transformations, and this has sometimes been the underlying rationale for thinking that a linear relation between two measures is an indicator of interval level interpretability of measurement results.

consequently be linearly related. Can we now treat them as mapping equalities and inequalities in differences in happiness? Surely not. All we know is that people are more inclined to endorse moderate rather than strong claims about their happiness. We have no idea what kinds of differences in happiness underlie different scorings, because the association of the observed scores does not manifest anything about the mixture of underlying factors (attribute, biases) that gave rise to the observed scores. This is true regardless of the details of how happiness *qua* mental attribute is construed (bearing in mind the commitment to non-operationally defined attributes).[71] We should therefore not use these single-item instruments to establish interval level hypotheses, such as the claim that "Having a child and taking a puppy increase happiness an equal amount".[72]

We have seen that neither coefficient alpha nor construct validation (in the present sense) produces evidence of quantitative representation. We have also seen that in both cases this failure is due to the fact that people's rating behaviour is influenced by a multitude of factors that go beyond the targeted attribute (such as rater effects). When the influence of these factors is unknown, we are unable to read people's ratings in terms of their standing on the attribute of interest.

If the problem with coefficient alpha and construct validation is that they have nothing to say about the black box of score determination, the obvious solution is to try to investigate that black box; in other words, to ask about the nature and composition of the factors that determine the way a person rates their standing on some attribute. IRT models do just that: they hypothesize factors, and relations between those factors, that might determine the probability of a given response on a rating instrument.

### *4.4.3 The Rasch model*[73]

In this section I will show that the simplest IRT model, the Rasch model, has a strong claim to being the best psychometric technique for ensuring quantitative

[71] In other words, I am *not* arguing that psychometricians have philosophically or otherwise indefensible conceptions of happiness, although that might be true as well.

[72] It is, furthermore, very rare that measures of the same or closely related target constructs would show a linear relationship in the construct validation exercise. This is especially true if the rating instruments differ with respect to the kind of judgment that the subject is asked to make, for example, whether they are asked to choose between categories or to make a comparative judgment (Guilford 1961, 115).

[73] Much of the material in this section can also be found in my forthcoming article on the complementarity of psychometrics and RTM (Vessonen 2018).

representation – at least in theory. I will show that the Rasch model is an instantiation of the *conjoint structure*, a structure that has been set forth within the RTM tradition and shown to ground interval level representation.[74] Being connected to the axiomatic foundations of interval scale properties in this way, there is a clear justification for why (analysis in terms of) the Rasch model can be thought of as a test of quantitative properties, in theory. The caveat "in theory" is important: the empirical goodness-of-fit tests between data and the model have many shortcomings, making it difficult to draw conclusive claims about quantitative representation in practice.

The theory of conjoint measurement (also known as additive conjoint measurement or simultaneous conjoint measurement) is one of the most celebrated axiomatizations in the RTM tradition. It was first proposed by psychologist R. Duncan Luce and statistician John Tukey in 1964. They were motivated by the fact that axiomatizations of quantitative measurement structures required that the attribute of interest allows side-by-side comparison and combination (i.e. concatenation) but psychological attributes do not allow for such operations. We can compare the length of rigid rods by observing them side-by-side, but we cannot set Andy's happiness next to Bobby's happiness and compare them. Luce and Tukey's proposed an axiomatization of conjoint measurement, which achieves interval level measurement of attributes that do not allow for side-by-side comparison or concatenation.

The axioms of conjoint measurement describe an empirical structure in which one attribute can be described as the simultaneous 'effect'[75] of two other attributes. More precisely, the axioms describe a situation where an attribute Y, which is the joint effect of component factors D and A, is the sum of the effect of A as captured by real-valued function $\phi$ and of the effect of D as captured by real-valued function $\psi$, i.e. $Y = \phi(D) + \psi(A)$ – in other words the component factors combine additively to form the joint effect. Luce and Tukey (1964) give the example that loudness of a tone can be thought as the effect of frequency and intensity of the tone.

[74] The interval property of the Rasch model has been proven also without reference to the conjoint structure. The conjoint measurement-Rasch connection is, in my view, the simplest way to explicate why Rasch instantiates interval level measurement. For other proofs, see e.g. Fischer and Molenaar (1995).
[75] I follow Luce and Tukey (1964) in using this term.

The Rasch model is another example of the kind of attribute structure conjoint measurement corresponds to, because it posits that the probability (or log-odds) of a correct response to a test item can be thought of as the effect of the difficulty[76] of the question and the ability of the respondent. The fact that Rasch model instantiates the additive attribute structure can be readily seen from the log-odds version of the Rasch model:

$$\ln \frac{(P_{is})}{(1 - P_{is})} = \theta_s - \beta_i$$

where

$P_{is}$ is the probability of a correct response to item i from subject s,

$\theta_s$ is the ability level of subject s, and

$\beta_i$ is the item difficulty parameter (see Embretson and Reise 2000, 148).

This mathematical connection between Rasch and conjoint measurement was first noted by Keats (1967) and since then several people have discussed it (e.g. Perline, Wright, and Wainer 1979; Andrich 1988; Wright and Stone 1999; Embretson and Reise 2000; Bond and Fox 2001; Borsboom and Mellenbergh 2004; Kyngdon 2008). However, the subject is not well-known and exists in the margins of the psychometric literature. It is also somewhat controversial as to what it means that the Rasch model 'instantiates' conjoint measurement.[77] It is hard to find a thorough explication of the relationship in the literature, which is why I shall supply one here.

For our purposes, the most important axioms of conjoint measurement are the *cancellation axioms*, and I shall focus on them.[78] To get a grasp of the cancellation axioms,

[76] The term "difficulty" is inherited from the educational settings, where Rasch is frequently applied and where the idea of "difficulty" is intuitive. In other fields, difficulty can be interpreted as "resistance to endorsing an item" or some such. For example, in measurement of openness, a more "difficult" item is one that people tend to be more resistant to endorse. For example, the item "I would tell a stranger my name" is less difficult than the item "I would tell a stranger my most shameful thoughts".

[77] Some people have disputed that the Rasch model instantiates conjoint measurement (for example Kyngdon 2008). I believe the disagreement stems from different readings of "instantiates", which is why a thorough explication is needed here.

[78] Other important axioms of the theory are called Weak Order, Solvability and the Archimedean axiom. My narrower focus is warranted, first, because the literature treats cancellation axioms as the crucial targets of empirical testing of the conjoint structure (Luce et al. 1991, ch. 21.8; Embretson and Reise 2000, 148-149). Second, it is possible that the Rasch model contributes evidence concerning the fulfilment of Solvability and the

consider Table 9, where each column corresponds to a value of one of the component factors (e.g. letters a, b, c are values for the component factor 'ability'), rows correspond to values of the other component factor (here letters d, e, f can be thought of as values for the component factor 'item difficulty') and the ordered pairs in each cell correspond to the joint effect (e.g. probability of correct response).[79] The so-called single cancellation axiom requires that the relative order of levels of the effect attribute for any two levels of one component factor is the same regardless of the level of the second component factor. Table 9 illustrates this axiom: the order of the cells for two values of the column variable (a and b) is the same for all levels of the row variable (d, e, f), as indicated by the sameness of the direction of the arrows:

| | a | b | c |
|---|---|---|---|
| d | (d, a) → | (d, b) | (d, c) |
| e | (e, a) → | (e, b) | (e, c) |
| f | (f, a) → | (f, b) | (f, c) |

*Table 9. Single cancellation. a, b, c are levels of one of the component attributes while d, e and f are levels of the other component attribute. The ordered pairs in the cells represent levels of the effect attribute.*[80]

Another important axiom, known as double cancellation axiom, is often expressed graphically as in Table 10. The verbal interpretation is that if the order relations indicated by the dashed arrows hold, then the order relation indicated by the solid arrow must also hold.

Archimedean axiom, but these contributions cannot be discussed independent of the details of the specific attributes under scrutiny (e.g. is the test measuring e.g. mathematical ability, quality of life etc.). This is because these axioms specify assumptions about the kinds of values each attribute can take. It is widely agreed that these axioms do not allow direct empirical testing but are rather accepted or rejected based on 'general considerations' (see Luce et al. 1991, section 21.8.4). Finally, the axiom of Weak order, involving conditions such as transitivity, is so weak that it is easy to see that the additive structure postulated in Rasch instantiates it.

[79] The presentation of the axioms differs slightly from article to article. I present the axioms in a form I take to be most common in contemporary literature (see e.g. Embretson and Reise 2000; Kyngdon 2008).

[80] To ensure that the tables are readable, not all arrows have been drawn.

| | a | b | c |
|---|---|---|---|
| d | (d, a) | (d, b) | (d, c) |
| e | (e, a) | (e, b) | (e, c) |
| f | (f, a) | (f, b) | (f, c) |

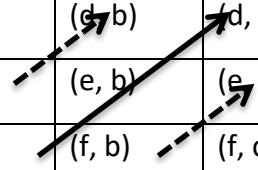

*Table 10. Double cancellation. Interpretation of columns, rows and cells as in Table 9.*

From the axiomatization that Luce and Tukey give, they arrive at their seminal conclusion concerning interval level representability:

*From the axioms we give, simultaneous measurement on interval scales is obtained for each kind of quantity separately and for their joint effects.* (Luce and Tukey 1964, 2)

In other words, their seminal representation and uniqueness theorems showed that if the axioms of conjoint measurement (only two of which have been presented here) are fulfilled, the component attributes as well as the effect attribute have a meaningful interval level representation. (For the full axiomatization and the theorems, see Luce and Tukey 1964).

With these axioms and representational results at hand, we get a more detailed account of the connection between conjoint measurement and the Rasch model. It is possible to illustrate the connection between the Rasch model and conjoint measurement by plugging in any item difficulty levels and any ability levels in the Rasch model and calculating probability values to complete a two-way table, as in Table 11 and Table 12.[81] It is easy to see that the above-described cancellation axioms are fulfilled when the Rasch model fits perfectly.[82]

[81] I am using the formulation of the Rasch model presented in chapter 2 because probabilities are easier to grasp than log-odds.

[82] Note that the ordering of values of the component variables (ability, item difficulty) is not relevant to a demonstration of the fulfilment of the axioms. For example, the diagonal arrows in table 12 can be drawn downward from left to right, and the values in the cells continue to fulfil the condition the arrows indicate. I have drawn only one set of arrows to ensure that the tables are readable.

| | Ability | | | | |
|---|---|---|---|---|---|
| | | **−1** | **0** | **1** | **1.5** |
| | **−1** | 0.50 | 0.73 | 0.88 | 0.92 |
| Item | **0** | 0.27 | 0.50 | 0.73 | 0.82 |
| Difficulty | **0.5** | 0.18 | 0.38 | 0.62 | 0.73 |
| | **1** | 0.12 | 0.27 | 0.50 | 0.62 |

*Table 11. Single cancellation and the Rasch model. The cells present probabilities calculated from the Rasch model. Interpretation of arrows as in Table 9.*

| | Ability | | | | |
|---|---|---|---|---|---|
| | | **−1** | **0** | **1** | **1.5** |
| | **−1** | 0.50 | 0.73 | 0.88 | 0.92 |
| Item | **0** | 0.27 | 0.50 | 0.73 | 0.82 |
| Difficulty | **0.5** | 0.18 | 0.38 | 0.62 | 0.73 |
| | **1** | 0.12 | 0.27 | 0.50 | 0.62 |

*Table 12. Double cancellation and the Rasch model. Interpretation of arrows as in Table 10.*

Tables 11 and 12 illustrate the meaning of the claim that the Rasch model instantiates the attribute structure of conjoint measurement: when the attributes have the structure postulated in the Rasch model, levels of the three attributes form the patterns postulated in the experimentally (dis)confirmable axioms. On grounds of Luce and Tukey's theorems, we also know that conjoint measurement yields an interval level representation. Following this, goodness-of-fit tests with the Rasch model can act as tests of whether a specific target attribute has the structure that allows it to be represented on an interval scale. How? Recall (from chapter 2) that the goodness-of-fit tests check whether the data conforms to the predictions of the Rasch model. As we see from Tables 11 and 12, the Rasch model predicts that the data forms the kinds of patterns that fulfil the cancellation axioms of conjoint measurement. So, the better the predictions of the Rasch model converge with patterns in

the data, the more we have evidence that the target attribute fulfils the cancellation axioms. In summary:

**Rasch1:** If we have evidence that the axioms of conjoint measurement are fulfilled, then we have evidence of an interval representation of the attributes of interest.
**Rach2:** If we have evidence that the attribute of interest has the structure postulated in the Rasch model, then we have evidence that manifestations of the attributes fulfil the axioms of conjoint measurement.
**Rasch3:** If empirical tests of fit between data and the Rasch model show that the data fits the model, then we have evidence that the attributes of interest have the structure postulated in the Rasch model. [83]
**Conclusion:** If the data fit the Rasch model, we have evidential support for interval representation.

Note that the truth of Rasch3 is trivial conceptually speaking, because the whole point of goodness-of-fit tests is to inform us about whether the attribute has the relevant structure. Empirically speaking, though, a good fit to the Rasch model does not clinch the case for the attributes having the postulated structure. There are a variety of reasons for this, for example, the procedures of model parameter estimation and the goodness-of-fit tests have their own shortcomings and implementation-related difficulties (Hambleton et al. 1991; Embretson and Reise 2000).[84] Sometimes the problem is not goodness-of-fit tests but the way the test is modified to produce data that fits the Rasch model. The fit of the model may be artificial, by which I mean that items have been dropped and modified until the data fits by construction. In such cases, the fit may be better characterized as a by-product of meddling with the test rather than evidence for interval representation of a non-operationally characterized target attribute.

These issues underline that a fit to Rasch model must be interpreted in terms of other evidence to justify the inference to adequate interval representation. As in other typical

[83] Assuming that the measure has been validated in other respects, e.g. that the data pertains to the attribute of interest (e.g. mathematical ability) rather than a different one (e.g. reading comprehension). Construct validation is the standard way of addressing this aspect.
[84] Kyngdon (2011) proposes tests that are meant to overcome blind spots of standard goodness-of-fit tests.

scientific endeavours, a fit to the Rasch model is just one piece of evidence that must be contrasted to other evidence to make the best possible inference. All this is to hammer the hopefully obvious point that Rasch, like any validation technique, is not all-encompassing.

Besides the issues with the implementation and interpretation of goodness-of-fit tests, there is an even more pressing practical obstacle to relying on the Rasch model in the quest to ensure quantitative properties. The problem is that the Rasch model rarely fits, for the obvious reason that more often than not the probability of a correct (specific) response is determined by more factors than just ability and difficulty. Even though the Rasch model was formulated some fifty years ago, it has not gained currency in mainstream psychometrics, plausibly because it is too demanding of a test of validity. The upshot is that the Rasch model is, theoretically-speaking, the best available technique for investigating interval properties, but in practice it tends to fail to deliver.

#### *4.4.4 Other IRT models*

We have seen two reasons for a failure to vouch for interval representation: i) black-boxing the interaction between observed responses and the target attribute, and ii) providing too simplistic of a model of that interaction. What about offering a more nuanced model of how the target attribute brings about observed response patterns?

That is exactly what more complicated IRT models are meant to do: they incorporate more parameters than just difficulty to account for the fact that response patterns rarely arise simply from the interaction of ability and difficulty. The excitingly-named 2PL model (referring to the two item-related parameters), for example, adds to the Rasch model *an item discrimination parameter*. The model has the following form:

$$P_i(\theta) = \frac{\exp\left(\alpha_i(\theta - \beta_i)\right)}{1 + \exp\left(\alpha_i(\theta - \beta_i)\right.}$$

where

$P_i(\theta)$ is the probability of a correct response to item i from a randomly selected examinee whose ability level is $\theta$,

$\beta_i$ is the item difficulty parameter, and

$\alpha_i$ is the item discrimination parameter.

The intuitive interpretation of the discrimination parameter is that it specifies how well each item manages to distinguish between the ability of subjects at different points of the ability scale. In other words, if an item is highly discriminating at some ability range, people with higher ability in that range are much more likely to give a correct response than people with lower ability. Put in yet another way, in this range the difference between an incorrect and correct answer is very informative of the subject's ability. A yet more complex model, the 3PL, incorporates a parameter for the influence of guessing. I will focus on 2PL here, but the conclusions generalize to more complex IRT models.

Unfortunately, the move to more complex IRT models is not a solution to the problem of establishing interval scales. The reason is that the item parameters that more complex models incorporate, although rendering the model more realistic, mess up a property that is crucial for making inferences to interval scales. This is the property that the difficulty ordering of the items should be the same across ability levels. That is, ordering of two items A and B in terms of difficulty must be the same for subjects of all ability levels. Let us call this property *difficulty ordering* (for more details, see Levine 1970; Cliff 1989; Embretson and Reise 2000).

We need a little bit more of the technical apparatus of IRT modelling to explain why difficulty ordering is important. Figure 6 depicts so called *Item Response Curves* (IRCs), which are a common tool for IRT users. The x-axis represents ability levels (estimated in the manner described in section 2.4) while the y-axis represents probabilities of correct response. Each curve corresponds to an item – hence the name *item* response curve. Points on each curve represent the probability that a subject of a given ability level gives a correct answer (or endorses the item, see section 2.4) on the item the curve corresponds to. For example, point A (indicated by the dashed lines) means that on the item the blue IRC corresponds to, the probability of correct response for a subject of ability level -1 is a little under 0.3.

The item response curves in Figure 6 have been calculated using the Rasch model. Consider the IRCs relative to each other. As suggested by mere eyeballing, their slopes are equal (as they would be for any other Rasch model IRCs). Due to their equal slopes, the IRCs do not cross. This, in turn, means that whatever ability level we look at, the ordering of the items in terms of their difficulty (that is, how likely a subject of that ability level is to pass

the item) is the same. In terms of Figure 6, if we look at, say ability levels -2 and 1, in both places "the orange item"[85] is the easiest (highest probability of passing) and "the grey item" is the most difficult.

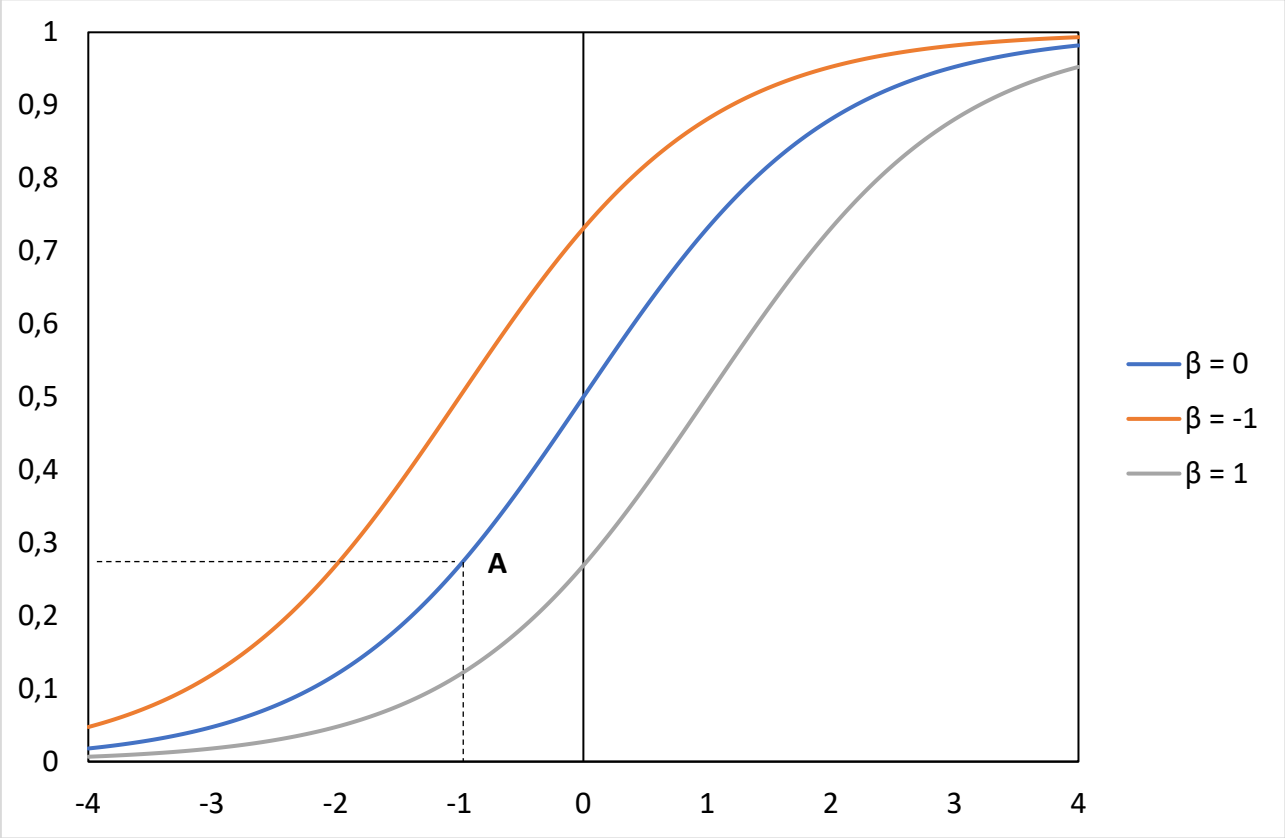


*Figure 6. Hypothetical Item Response Curves for the Rasch model. Item difficulties are depicted on the right.*

Now look at Figure 7, which depicts IRCs for the more complex IRT model, the 2PL. The IRCs have different slopes, which leads them to cross. This means that the ordering of the items in terms of difficulty varies across ability levels. For subjects with ability 2 "the blue item" is the easiest and "the grey item" the most difficult, but for subjects with ability -1 the ordering is reversed. Many authors regard such crossing as a death sentence to interval representation (Cliff 1989).

[85] That is, the item represented by the orange IRC.

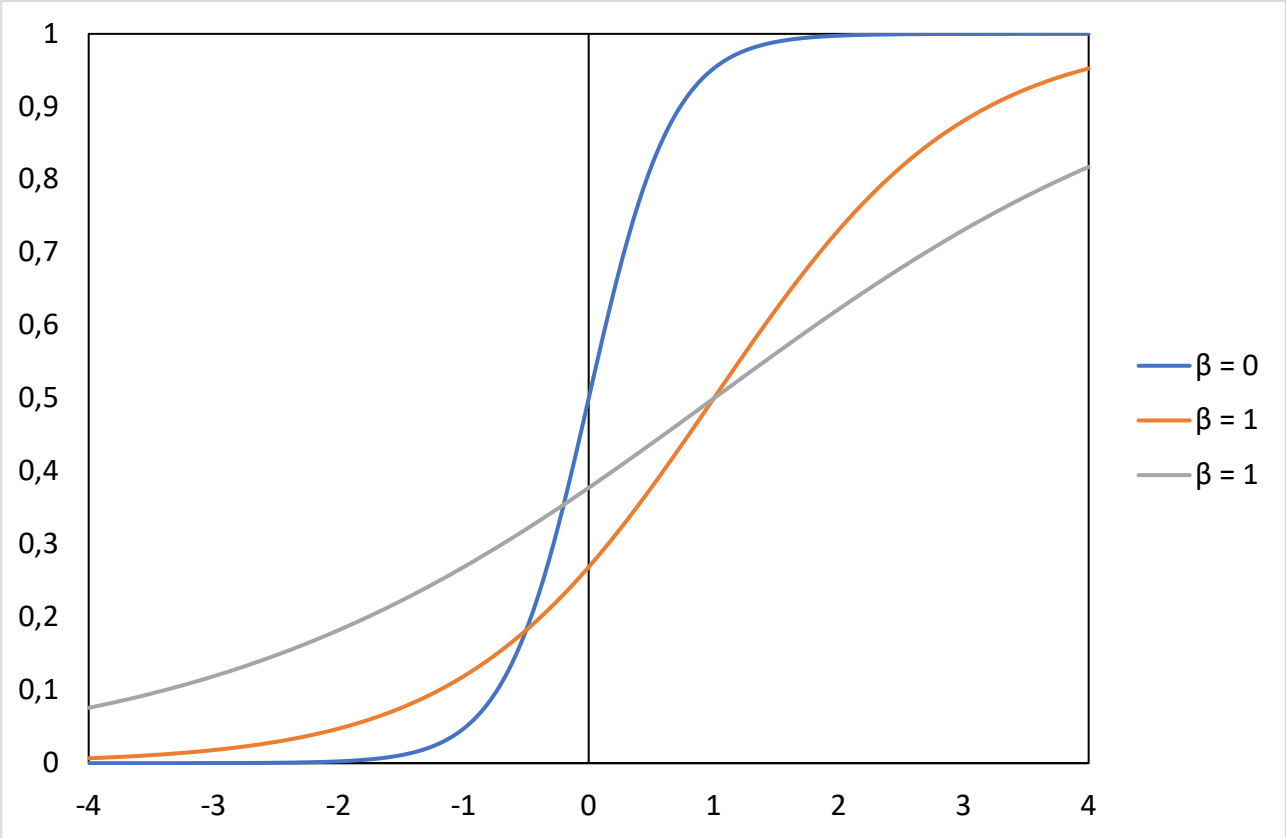


*Figure 7. Hypothetical Item Response Curves for the 2PL. Item difficulties are depicted on the right.*

In non-technical terms, the problem with crossing item curves is that intervals on the ability scale do not have a constant interpretation in terms of how people (are likely to) respond in the test as a whole (and on further extensions of the test). More precisely, different intervals on the ability scale mean different things in terms of a person's ability to give the correct response on one item relative to their ability to give a correct response on another item. For example, in Figure 7, intervals above the zero point (and slightly below since the crossing is not exactly at zero) of the ability scale entail higher probability of success on the blue item than on the grey item, while the opposite is true for ability levels below zero. With more items, and more crossing IRCs, the interpretability of intervals in terms of what people can do becomes more and more opaque. Such variance of interpretations of intervals does not plague the Rasch model, because the IRCs do not cross.

A defender of complex IRT models might respond: it is not complex at all to make inferences about what the intervals mean – just look at the item curves at the ability intervals of interest and read the meaning! Sure, she might continue, it can get a little laborious, but

the information is there for anyone who is interested. In response, I would like to draw attention to what I take to be the likely source of crossing in the psychometric context. As said, crossing IRCs entail that equal differences on the ability scale mean different things in terms of what people can do. In fact, equal differences on the ability scale mean (in a sense) *opposing* things in terms of what people can do – some intervals mean that the grey item is harder than the blue one, and *vice versa* for other intervals. Why might the same ability difference lead to various, even opposing behavioural consequences? Plausibly because the intervals represent attributes that are only nominally "the same ability". That is, although the ability scale is treated as unified when it is presented on the x-axis of the IRC graph, different intervals on that scale represent (slightly) different attributes, each of which drives different kinds of behaviours. The blue IRC might, for example, represent a question involving a mathematical operation that children learn at a specific age (say, multiplication). The item therefore distinguishes sharply between children below that age and children above that age (hence the steep slope of the IRC). The grey item, by contrast, might correspond to a task that children get gradually better at throughout childhood. Different slopes of the IRCs (and therefore their crossing) are explained by the fact that each item represents a different (but potentially related) attribute. [86] If this is so, it is hard to see how the test as a whole could be said to represent a single target attribute on an interval scale – unless we go operationalist about the attribute.

The above-described problem is a variant of an observation Geoffrey N. Masters, now CEO of the Australian Council for Educational Research, made some 30 years ago in *the Journal of Educational Measurement* in an article entitled "Item Discrimination: When more is worse". (Masters 1988). Masters argued that in educational settings, item discrimination (the parameter that drives the crossing of the curves in 2PL) may be due to test bias. For example, one item may discriminate the way it does because it measures a combination of ability and proneness to fatigue and frustration, while another item measures a combination

[86] One might object that the same attribute might manifest in different ways in different kinds of tasks. For example, while higher intelligence usually leads to better success, in some tasks higher intelligence might lead to worse outcomes, because highly intelligent people think of the problem in too complex ways. As discussed in chapter 3, delineating attributes is difficult. My point here is that the crossing of the item curves can at least sometimes be explained by unintended measurement of two different attributes, which in turn threatens claims about interval level representation. The fact that crossing is not always thus explained does not undermine my point.

of ability and what Masters' calls "test-wiseness", i.e. the ability to give the correct answer based on knowledge about testing, not knowledge about what is being tested. One way to recast this point is to say that the items measure different target attributes, each of which is composed of an ability aspect and a bias aspect. Whether or not one likes this re-description, the upshot is the same: the inference to meaningful interval level representation of a non-operationally characterized attribute is disrupted.

The problems I have outlined may sound like overconcerned "ifs". But the examples illustrate a relevant point: a fit to the complex IRT models in and of itself does not ensure interval level representation. Just like in the Rasch case, the fit must be interpreted in terms of other evidence and theoretical considerations to warrant the leap to claims about quantitative representation of a non-operationally characterized target attribute. Papers on the validity of psychometric instruments sometimes interpret IRT analyses vis-à-vis other evidence and theoretical considerations, but the papers rarely use such interpretations to make inferences about quantitative representation.[87] In the next section, I will provide a sketch of how to justify claims about quantitative representation in psychometrics.

### 4.5 What now?

I have argued in favour of the following two premises:

P1. Psychometricians assume an interval representation of a non-operationally defined target attribute.

P2. Psychometric validation usually fails to validate an interval representation of a non-operationally defined target attribute

The upshot is that:

C. In psychometrics, the assumed representation is usually not validated.

[87] For example, IRT analysis is often used for detecting whether items behave differently in dissimilar populations (e.g. is there something about the item that leads women to receive lower scores even when they have the same ability as men). This is called analysis of DIF or differential item functioning in the technical jargon. Researchers frequently attempt a verbal explanation of DIF in the discussion section of their paper (e.g. Oishi (2006) on life satisfaction) but do not explain how DIF inhibits a quantitative interpretation of the results of the instrument.

This conclusion matters, because as saw in the very beginning of this chapter, representational assumptions that have not been validated can hinder our ability to interpret psychometric instruments. For example, when changes in average scores on a depression scale fail to inform us about changes in depression, we are also likely to fail to establish truths about the efficacy of a new drug.

What to do about the mismatch between representational assumptions and validation? Reacting to old representational challenges, some authors suggest that psychometricians should accept that interval scales are unattainable and move to ordinal test theory (see Cliff 1989 for an overview of attempts). The downside of this solution is obvious: many of the typical operations in psychometricians statistical toolbox make no sense with ordinal data. The Rasch enthusiasts, by contrast, argue that tests need to be modified until they fit the Rasch model. The downside is that discarding items until the Rasch model fits will typically result in impoverished tests. Others argue that if we study rating behaviour meticulously, we can standardize the meaning of intervals (see e.g. Veenhoven 2009). Another possible solution is to try to achieve representation on scale types that do not occur in Stevens' classic categorization. Such endeavours, while intriguing, require technical prowess and willingness to work in the margins of psychometrics, which many arguably do not have. Finally, one could give up the idea that psychometric instruments need to represent non-operationally characterized attributes. Perhaps psychometrics is in the business of assessment of test behaviour, not quantitative representation of a test-independent reality? The difficulty is to show that psychometric instruments are useful despite not representing non-operational attributes.

I think it is useful to divide the above-mentioned strategies according to the premise they pertain to, that is, whether they tackle Premise 1 concerning quantitative representational assumptions, or Premise 2 concerning the adequacy of validation methods to ensure the truth of representational assumptions. The strategies listed above fall rather neatly into groups according to the premise they try to falsify – Table 13 provides my attempt at such a classification. In my view all these efforts are valuable, because they constitute attempts to make the psychometric practice consistent vis-à-vis representational questions.

| Letting go of the assumption of quantitative representation | Establishing quantitative properties |
|---|---|
| Move to ordinal test theory | Rasch approach |
| Broaden the classification of scales | Better handling of bias |
| Move to classification | |
| Give up representation of non-operationally defined attributes | |

*Table 13. Ways of dealing with the New Representational Challenge.*[88]

Before concluding, I would like to offer some of my own thoughts about how to resolve the inconsistency. My focus will be on the validation of the quantitative representation of non-operationally characterized attributes, that is, Premise 2. Chapter 6 will be more concerned with Premise 1.

My first suggestion is that the validation process needs to pay due attention to the similarity condition outlined in chapter 3, that is, the requirement that the target attribute must be similar across the empirical relational structure that is being numerically represented. In chapter 3, my example was that it ought not to be the case that, for example, the ordering of numerals in one part of the numerical representation represents strict ordering in terms of weight, while in another part the ordering of the numerals represents strict ordering in terms of height. The reason is that the interpretability of the numbers suffers from such "mixedness" of representation – the numerals would have to be flagged for the attribute they pertain to in order to be correctly interpreted. Interpretability suffers particularly much if a single ordered pair of numerals represents a combination of multiple attributes, for example, if the numerals represent the sum of a subjects' weight and the width of their lap. The problem with such a representation (or one of the problems) is that equality of assigned numerals cannot be interpreted as equality in terms of the constituent attributes: for example, in a strict ordering representation of such a mixed empirical relational structure, Rei (140cm and 30kg) and Jalo (135cm and 35kg) would be assigned the same numeral, but clearly their status on the constituent attributes is different. This is especially problematic when the mixed attribute has no shared or intuitive meaning, such as in the case of the combination of one's weight and the width of one's lap.

The difficulty of interpreting representations of mixed attributes has not gone unnoticed in psychometrics. In the literature, these concerns are expressed, for example, in

[88] In this table, rows do not have an interpretation.

terms of the need for attribute homogeneity, unidimensionality and first-factor-saturation (Thomson 1940; Loevinger 1947). The proposal to focus on similarity is, in other words, not a novelty, but I do think it deserves more attention from those psychometricians whose genuine aim is to validate quantitative representational assumptions. For instance, if our measure of depression loads on four factors (i.e. a reasonably well-fitting model has four factors) – say, factors characterized as anxiety-related, depressive mood-related, insomnia-related and somatic symptoms-related (Shafer 2006) – this is a fairly sure sign that interval level interpretability of the total score is endangered (because the total score masks divergent constituents). Nonetheless, such multidimensional instruments are commonly treated as if they yield interval level representations – as is the case when, for example HAM-D is used to make quantitative claims about change in depression (see sections 4.2 and 4.3 above). In sum, my first recommendation is to *focus more on justifying judgments of homogeneity.*

The second, perhaps more contentious recommendation pertains to interpretations of IRT model-fitting exercises. I have argued that in and of itself, an IRT analysis does not allow psychometricians to infer quantitative representation. What is crucial is how one interprets patterns of fits and non-fits in IRT analyses, how one juxtaposes these to other evidence, and what one does about the outcome. [89]

What would this look like, in practice? Consider an implausible but illustrative example: one ability test item discriminates sharply between candidates (compared to other items!), because the answer to that question was accidentally left on the wall of the class room, and only half of the subjects were able to read the answer due to their advantageous position in the room. The crossing of this item curve and others is clearly due to undesirable bias, not the attribute of interest. If the researchers have knowledge about the test set-up, or at least evidence of low test-retest reliability for this item, they can infer the poor quality of the item and proceed to remove it from the analysis. This would enhance their ability to make inferences about quantitative representation.

Consider another example: it turns out that items on a depression test can be categorized roughly to two groups, one of which has highly discriminating items while the other has low-discriminating items. Suppose further that reports of the interviewers who administered the test suggest an explanation: the highly discriminating items seem to evoke

[89] Thanks to Eran Tal for discussions that helped me make this point clearer.

an exaggeration reaction in some of the subjects, plausibly because the questions involved "subjective" questions about feelings rather than objective questions about bodily functions. This hypothetical explanation prompts researchers to study the subjects' response styles further. It turns out that the subjects suspected of exaggeration respond similarly over time – their exaggerative tendency persists. Other subjects receive lower scores on retest – suppose expert psychiatrists say that such a change corresponds to expected variations in depression. This evidence can be used to adjust depression test responses according to proneness to exaggeration – this would again increase the plausibility of claims about interval scale representation of the target attribute.

These examples are purposefully simplistic. What they illustrate is that to infer quantitative representation, what matters is juxtaposition of results from different tests, and their rich interpretation *from the perspective of quantitative representation.* In other words, it is not enough to know the results of various statistical tests – psychometricians are already incredibly good at the technical analysis of patterns in the data. For the tests to justify claims about quantitative representation, the technical results need to be interpreted and handled specifically in terms of their bearing on quantification. This recommendation may be old news to some psychometricians, but as I summarized in the beginning of section 4.4, I believe such cross-interpretation in terms of quantitative representation is not a common practice in psychometrics. Hence my second recommendation for those in search of quantitative representation is this: *juxtapose results from various validation techniques and evaluate their juxtaposition in terms of quantitative representation.*

### 4.6 From representation to the represented

In this chapter I have constructed an argument I call the New Representational Challenge, which concludes that in psychometrics, the assumed representation of target attributes is usually not validated. In addition, I outlined some possible avenues for rendering psychometrics consistent again.

Thus far this chapter, and the dissertation as a whole, has focused on representation at the expense of what is being represented. I have argued that it is reasonable to think that psychometricians often try to represent non-operationally characterized attributes, but I have not spoken much about what that entails and whether that is the only

viable option. In the next two chapters we will reverse the focus: representation takes the place of a passenger, while that which is being represented – attributes and concepts – take the driver's seat.

# 5. Conceptual Engineering

> *[T]he commitment one made early on in life to a particular cutting up of the world at its joints is hard to see as merely one possible commitment among many, and just as it is hard to see, it is hard to let go of.* (Bergen 2012, 194)

## 5.1 Controversial concepts

In philosophy of science, concept formation – the activity of characterizing and defining concepts for the purposes of scientific or everyday usage – receives less attention than hypothesis testing. Nonetheless, at least two long-standing philosophical debates revolve, at least implicitly, around concept formation. One of these debates involves questions about the relationship between the natural and the constructed in the formation of adequate scientific concepts.[90] I call this debate *the construction debate.* The other debate concerns intuitions: to what extent should our conceptual apparatuses mirror folk intuitions or the intuitions of philosophical experts? I call this *the intuition-pump debate*.

Let's start the tour at the side of construction. I take it that everybody agrees that definitions are created by humans – we do not find definitions lying around in nature, we make them. In a trivial sense, then, scientific concepts are constructed. But is there more construction involved than that? Many philosophers, who sometimes call themselves naturalists, think that that is where construction ends. According to naturalists, our conceptual apparatuses ought to cut around nature's joints, that is, track natural as opposed to human-made structures. The periodic table of chemical elements is often used as an example of successfully cutting around nature's joints (e.g. in their encyclopaedia entry on natural kinds, Bird and Tobin (2018) use this example ubiquitously). On this view, concepts are discovered, not constructed.

[90] What realness and naturalness means in the context of concept formation is slippery. By considerations over reality I shall mean investigations (usually empirical in nature) to the existence and characteristics of entities that are or are thought to be in the denotation of a concept or a concept-to-be. More generally, reality denotes that which we treat as reality when we are investigating matters of interest through our best and most appropriate means of sensing and experiencing, whether scientific or otherwise. The conceptual points I present in this chapter should be easy to digest without a strict theory of realness or naturalness.

Naturalism contrasts with conventionalism,[91] which comes in a few different stripes (e.g. Bird and Tobin 2018). Its core claim is that scientific concepts are more human-dependent than naturalists take them to be. In radical forms, conventionalism states that all similarities and differences that underwrite our classifications are human-dependent. For example, the similarities and differences we use to assign zebras and horses to different classes are, on this extreme view, somehow contingent on human activity and discourse.[92] In less radical forms, conventionalism insists that *some* things that appear human-independent and natural may in fact be human-dependent, that is, brought about or shaped by social institutions and norms (see Hacking 1999). Gender and race are frequently cited as examples of such seemingly natural categories that turn out to be contingent on human practices on closer inspection (e.g. Haslanger 2000, 2008).

Naturalism and conventionalism are two opposing views in multiple on-going debates. Sometimes these debates concern specific concepts, for example, when scholars are arguing about the relative merits of naturalist and conventionalist analyses of concepts such as disease (Stegenga 2018), race and gender (Haslanger 2008). At other times, naturalism and conventionalism are positions in debates about concepts in general (e.g. Brigandt 2011; Ereshefsky 2018 for recent contributions). In sum, the matter of concern is the balance of the natural and the constructed in the formation of characterizations of concepts. This is the essence of what I call the construction debate.

Turn to the intuition-pumping debate. Everybody knows that one of philosophers' main tools is the construction of thought experiments and the analysis of the intuitions those experiments evoke.[93] When thought experimentation is applied to concept formation, the structure of the exercise is typically the following: i) definition D of concept C is proposed, ii) thought experiment T involving concept C is constructed, and iii) D is either accepted or rejected as an adequate definition of C, depending on whether intuitions regarding the application of C in T match definition D. As examples, think of Nozick's

[91] This is sometimes called "constructionism" as well.

[92] In their encyclopaedia entry, Bird and Tobin (2018) seem to assign this radical view to Woolgar (1988). They give no other references as to who upholds this view, but I follow their example in introducing this extreme conventionalism as one alternative that is on the table.

[93] Cappelen (2012) offers an extensive objection to the common wisdom. But given my broadly applicable notion of intuitions in section 5.5 below, I (echoing Chalmers 2014) believe the common wisdom still stands despite Cappelen's objections, which rest on a narrower conception of intuitions.

experience machine or Goldman's fake barns, and how these thought experiments have been used to argue against certain definitions of well-being and knowledge, respectively.[94] The activity is ubiquitous, but increasingly debated (Shepherd and Justus 2015). What is the point of intuition pumping? Do intuitions allow access to conceptual truths? Need concepts be maximally intuitive? What if intuitions disagree? Are philosophers' intuitions better than folk intuitions? These and similar questions form what I call the intuition-pumping debate.

In this chapter I define and defend a specific approach to concept formation, *conceptual engineering*, and show how it navigates these debates. Although conceptual engineering has become popular among philosophers and methodologists in recent years – with publications, conferences and research projects dedicated to exploring it[95] – there is no unified definition or framework for its application. I shall therefore define and defend my own version of conceptual engineering in this chapter. An explication of conceptual engineering is needed for the purposes of chapter 6, where I articulate my view of how psychometric validation can be used to engineer different types of concepts.

In section 5.2, I provide a characterization of concepts. Building on this notion of concepts, I outline conceptual engineering in section 5.3. Section 5.4 defends conceptual engineering against the charge that it disregards reality, while section 5.5 explicates the relationship between conceptual engineering and conceptual analysis. Section 5.6 concludes.

## 5.2 What are concepts?

### *5.2.1 Minimal criteria*

Before I can put forth a methodology of concept formation, it is useful to think about what exactly it is we are forming, when we are forming concepts. Are we forming mental

[94] There are various versions of the experience machine thought experiment (originally formulated by Nozick 1974), but the core idea is the following. Imagine a person that is attached to a machine that produces lots of pleasurable experiences. If well-being is mere pleasurable experiences, such a person is very well off. But that seems counterintuitive. Hence, well-being must be something else. Gettier style examples are meant to show that "justified true belief" is not an adequate definition of knowledge. The standard example (by Goldman 1976) goes like this: imagine you are driving in a field and you see what looks like a barn and belief that you see a barn. And in fact, it really is a barn you see. It seems that your observation has led you to form a true justified belief. What you don't know is that the area you are in consists primarily of fake barns or barn facades that look like barns but really aren't. Some people intuit that the person seeing the barn does not really have knowledge.

[95] The ConceptLab at the University of Oslo is dedicated to studying conceptual engineering. The first book length treatment of the topic is by Herman Cappelen (2018). A 2018 conference at the New York University had the title "Foundations of Conceptual Engineering", and conferences and workshops with similar themes have been or will be organized in at least the Catholic University of Louvain, the University of Toronto and the University of Zürich.

representations, writing necessary and sufficient conditions, reaching out to Platonic forms, or something else entirely? That, it turns out, is not such an easy question to answer to. The nature of concepts has been debated basically throughout the history of philosophy (see contributions to Margolis and Laurence 1999). Here is a taste of these debates:

- One debate concerns the ontology of concepts: are concepts mental representations, abstract objects, or something else entirely (Sutton 2004; Margolis and Laurence 2007)? On the former approach, concepts are in the mind of their holder. For example, the concept of a dog is a representation of a dog in the mind of the person who knows the concept of a dog. On the latter approach – which has its origins in the work of Gottlob Frege (e.g. 1892) – concepts are not in the mind. Instead, they are abstract objects of some sort: they do not exist in a particular place and they exist whether or not anyone holds the concept. The concept of a dog, for example, is an *abstract* entity that is the *object* of people's thought when they hold beliefs about dogs.

- Another debate concerns the structure of concepts (for an overview, see Gerring 1999; Margolis and Laurence 1999). Are concepts structured like definitions, that is, in terms of necessary and sufficient conditions? Or is their structure their embedding in a theory, that is, the meaning of a concept emerges from its connections to other concepts in a theory? Or perhaps the Wittgensteinian family resemblance approach is the way to go: the things that fall under the concept form "a complicated network of similarities over-lapping and criss-crossing" (Wittgenstein 2009, 171)[96].

- There is even debate about whether there are concepts. Machery (2009), for example, defends what he calls concept eliminativism, according to which there are no concepts. Machery's argument starts from the claim that useful scientific notions are natural kinds – according to Machery, a natural kind is, roughly, a class where the referents have a large cluster of properties in common and these commonalities are the result of the working of a shared causal mechanism. Machery then argues that

[96] Originally published in 1953.

concepts, as they are used in contemporary psychology, are not natural kinds. Hence the concept of concept ought to be eliminated from contemporary psychology, according to Machery.

Do we need to take a stance on all these debates (and then some) to make headway in the methodology of concept formation? I don't think so. What we need is a minimal characterization of concepts – Concept Minimalism as I shall call it. Analogous to Representation Minimalism, Concept Minimalism is maximally neutral with respect to controversies about concepts. Neutrality with respect to on-going controversies means that when we build a method of concept formation on Minimalism, stakeholders with different, even opposing theoretical commitments can make use of the method.

Here is how I define Concept Minimalism:

***Concept Minimalism.*** *Concepts are building blocks of thoughts that can be characterized from at least two perspectives, namely, their intension (connotation) and extension (denotation).*[97]

And here is what Minimalism entails, more elaborately:

- Concepts are entities that occur in thought and that can be expressed, presented or represented by standard means of expressing thoughts, for example, speech, writing and visual arts.[98]

- Concepts can be characterized with reference to their extension, that is, the kinds of entities, phenomena and attributes they are meant to apply to (Gerring 1999; Goertz 2006). The concept "dog", for example, is usually meant to apply to four-legged

[97] As elaborated on below, this notion builds on concept formation literature in social sciences and philosophy. I say "at least" two perspectives, because there is fascinating research that suggests there might be other perspectives as well. For example, research on *embodied simulation* suggests that one important part of concept meaning is the cognitive process that starts when a person encounters (the term associated with) the concept. I think this line of thinking about concept meaning is compatible with the present approach, but I do not have the space to elaborate on this point here. Note also that by the term "building block" I do not mean to imply that concepts remain unchanged when they connect with other concepts to form a thought.

[98] Not all concepts can be *fully* captured by these means, but I take it that at least sketches are usually possible.

domesticated animals that descend from wolves. The concept “warmer than boiling water” typically applies to things that either feel or measure as warmer than water that has been heated to 100 degrees Celsius. And so on and so forth. The extension may also be empty, and in such cases we might say that the concept has potential to refer to entities (Goertz 2006). The extension may contain non-concrete entities, like ideas – as is the case with the concept of a concept itself.

- Concepts can also be characterized with reference to intension, that is, the way in which the concept presents things that are in its extension (Gerring 1999; Goertz 2006). This is the “meaning” or “sense” of a concept, because it specifies what aspects of the extension the concept brings forth. The stock example (due to Frege (1982)) is that the expressions “Evening star” and “Morning star” have the same extension, the planet Venus, but they present different aspects of their common denotation. Similarly, the extensions of concepts like “poison” and “medicine” can overlap partially, but clearly that common extension is presented in different ways by the two concepts.

- Sometimes the extension or intension of a concept is difficult to characterize in full, such as with contested or multifaceted concepts like “intelligent”, “god” and “disease”. But even in such cases incomplete sketches of the extension and intension are usually possible. For example, there are clear cases that everyone familiar with the concept of disease would assign to its extension.

I think that Concept Minimalism is neutral with respect to many of the controversies that characterize philosophical work on concepts. Here is how:

- Firstly, concept minimalism is open to various ontologies of concepts. Both mental representations and abstract objects can be described from the perspectives of extension and intension. Indeed, the “abstract object” -approach is grounded in the work of Frege, who is also credited as the originator of the intension-extension distinction and its application to concepts. Mental representations can also be

described from the perspectives of *how* they represent (intension) and the entities that *fall in the purview of the representation* (extension) (Margolis and Laurence 2007 present an ontological view that concords).

- Secondly, Concept Minimalism is neutral with respect to structures, that is, Concept Minimalism sets no restrictions on appropriate concept structure. In other words, the Concept Minimalist grants that the intension of a concept (and the extension thus implied) can be characterized in terms of necessary and sufficient conditions, family resemblances or the concept's place in a theory.

- Third, and finally, I don't think that even Machery's eliminativism threatens Concept Minimalism. The empirical claim that concepts *qua* natural kinds do not exist may well be true. But it is not true that the concept of a concept must be a natural kind in order for it to be useful.[99] To see this, consider first the fact that science is filled with notions that have *empty* denotations, from frictionless planes to perfectly rational agents. If a concept captures nothing at all, it certainly does not capture things that have a large cluster of properties in common, where these properties are brought about by a common causal mechanism.[100] But one would be hard-pressed to say that frictionless plane and rational agent are useless notions.[101] This suggests that the natural kinds-based criterion of usefulness is not always apt, which in turn entails that Concept Minimalism can well be a useful account of concepts.[102]

We now have an account of what concepts are, at the minimum. Thus, we know, roughly, what it is that a methodology of concept formation should guide us to form. But

[99] Machery (2009, 8.2) himself leaves for example folk notions out of the purview of his criterion of usefulness, that is, folk notions need not be natural kinds to be useful.
[100] Recall that this was Machery's criterion of something being a natural kind.
[101] One might of course say that frictionless plane and rational agent are not concepts but something else, but I think this would be a radical departure from how the concept of a concept is typically used.
[102] This might look like a rejection of Machery's view, rather an indication that Concept Minimalism is compatible with Machery's eliminativism. The way I read Machery is that his argument pertains specifically to the concept of a concept as it is used in *contemporary psychology*. If this is so, then it is possible that he would accept Concept Minimalism as a useful account of concepts for other contexts.

before we get to the real deal, the formation of concepts, I want to briefly set forth how the just outlined concept of a concept relates to attributes.

### *5.2.2 Concepts, attributes, entities – some technicalities*

The literature on concept formation and measurement is ridden with notions that sound similar to the concept of a concept just outlined. Examples of such similar-sounding notions are qualities, properties, attributes, quantities and features. The rest of this chapter will be easier to digest if we clarify how the above-described concept of a concept relates to these other technical notions, in particular, how concepts relate to attributes.

Let's start by distinguishing two types of things that can fall in the extension of a concept. On the one hand, there are concepts that refer to entities or objects: the concept of a dog refers (roughly) to four-legged domesticated animals that descend from wolves, while the concept of a chair refers (roughly) to objects that have the function of supporting a sitting person. Some concepts, by contrast, refer to aspects of those entities: for example, the concept "furry" refers to a quality certain dogs have, while the concept "comfortable" refers to a feature only some chairs have. Referents of this latter type are what I call attributes or properties. (As one can see, the distinction between these two types of concepts resembles the division of words to nouns and adjectives.)

What is important is that attributes are indeed *referents* of certain concepts, not a type of concept. The concept "comfortable" is an idea that can be verbally defined, while comfortableness as an attribute is an aspect of an entity. Just like the concept of a dog, or the definition of a dog, is not itself a dog, so the concept of comfortableness is not comfortableness itself. Similarly, in the measurement context, the concepts of length, intelligence and happiness are verbally expressible ideas that have the corresponding attributes as their extension. In my terminology, then, attributes are the same or very close to qualities and properties: they are what certain entities bear, possess or exemplify.[103]

For measurement, the consequence of the just-oultined distinction between concepts and attributes is the following. A measure measures (or tracks, or captures) an aspect of an entity, that is, an attribute. For example, a thermometer measures an attribute

[103] I shall try to not commit ourselves to any specific theory of properties and what it means for an entity to have a certain property. For a summary on rival views of the nature of properties, see Orilia and Swoyer (2017).

of bodies, and we usually call that attribute temperature. A thermometer does not measure a *concept* of temperature, strictly speaking, because a concept of temperature is an idea, not an attribute.

Nonetheless, it is often useful to talk about "the measurement of a concept of temperature" as short hand for "the measurement of an attribute that is in the extension of a specific concept of temperature". This is because the concept of temperature is used in different ways in different contexts, and those different variants of the concept have different (although usually partly overlapping) attributes in their extensions. For example, in many everyday contexts we think of the concept of temperature in phenomenological terms, that is, in terms of how hot or cold objects *feel* to us. Weather forecasts sometimes label this "feels like temperature". But "feels like temperature" is not the same as the physical scientists' concept of temperature, which is defined e.g. in terms of mean kinetic energy of molecules. The two concepts of temperature have different attributes in their extensions (a sensation versus a physical property), and hence their associated measures can rightly give divergent readings in the same circumstances. (In the weather forecast, the predicted "feels like temperature" frequently diverges from the predicted temperature.) When I speak about the measurement of a concept of temperature, I mean measurement of an attribute that is in the extension of some concept of temperature. This leaves room for a plurality of related conceptions of temperature. This mode of speaking also leaves room for discussing and re-shaping the extension of a particular concept – in other words, it leaves room for the act of concept formation.

Having now defined how my notion of concepts relates to attributes, we are ready to move to the real subject of this chapter: the formation of concepts. In minimalist terms, the question we are up against is: "How should we determine the intension and extension of concept X?" where X could be well-being, intelligence, temperature, disease or any other concept of interest to a measurer.

## 5.3 How to form concepts?

### *5.3.1 Carnap the conceptual engineer*

Does the new drug harm its takers? Does reading novels increase well-being? To answer such questions, some preliminary characterization of the concepts "harm" and "well-

being" is needed, if only to know what observations to treat as *potential* instances of harm and increased well-being. Frequently, such characterizations build on pre-existing conceptual apparatuses. For example, a psychologist working on well-being should probably respect *some* of the implicit intension and extension that manifests in common usage of the concept "well-being". Why? The psychologists will have a hard time justifying the term "well-being", if the corresponding concept in no way resonates with laypeople's conceptions of well-being. Why call it "well-being" if there is no continuity with what laypeople think well-being is? Similar considerations apply to specialist circles. For instance, the psychologist has good reason to align her conception of well-being with prominent psychological, sociological or even philosophical theories of well-being. Following Rudolf Carnap (1950b), I will call it *explication* when a new, explicit characterization of a concept is formed based on prior usage.[104]

Explication, as Carnap and his many followers treated it, is a method of *conceptual engineering* (Scharp 2013; Eklund 2015; French 2015; Brun 2016; Cappelen 2018).[105] For conceptual engineers, successful concepts are thoughtful answers to questions of usefulness. To put it in another way, in conceptual engineering the acceptability of a characterization of a concept depends on its appropriateness to the situation in which the concept has been or will be used. Practically, this means that a plethora of evaluative criteria may be used to assess a characterization, from "similarity to everyday usage" to "novelty" and from "correspondence to philosophical intuitions" to "simplicity". The criteria and their sources vary, as we will see, which makes conceptual engineering open-ended and almost trivially widely applicable.

What characterizes conceptual engineering is tolerance, an idea that occurred famously and prominently in Carnap's later work on language systems (e.g. 1950a). He wrote:

> *Let us grant to those who work in any special field of investigation the freedom to use any form of expression which seems useful to them; the work in the field will*

[104] Carnap (1950a) was motivated by the question: how are successful new concepts created in science? While Carnap's examples came from natural sciences (e.g. fish, temperature), for our purposes let us consider science that deals, for example, with concepts such as "disease", "intelligence", "harm" and "well-being".

[105] I shall frame conceptual engineering with respect to Carnap's philosophy, but similar emphasis on usefulness is central to pragmatist approaches to concept formation. A recent pragmatist account of concepts is due to Brigandt (2011). The conceptual engineering approach also bears resemblance to conceptual ethics (Burgess and Plunkett 2013) and ameliorative analysis (Haslanger 2012). See also Dutilh Novaes (2018) for a comparison of conceptual engineering and ameliorative analysis.

> *sooner or later lead to the elimination of those forms which have no useful function. Let us be cautious in making assertions and critical in examining them,* ***but tolerant in permitting linguistic forms.*** (Carnap 1950a, 40, emphasis original.)

Tolerance applies to conceptual systems as well as the criteria that guide the formation of concepts. For example, it would not be tolerant to adopt an *a priori* conviction, according to which acceptable concepts correspond to so-called natural kinds, that is, roughly, groupings that reflect (allegedly) natural structures. Concepts that capture groupings such as "Things that are green and observed before time *t*, and blue otherwise", "Dogs I'd name 'Marshall Woof'" and "Regularly occurring clusters of psychological symptoms with no known biological cause" are good concepts, if they serve a purpose. What might those purposes be? Making a philosophical point about induction (Goodman 1954), choosing a puppy, or helping patients with debilitating but complex mental problems, for example. Broadly speaking, then, conceptual engineering only rules out one thing: attempts to justify a choice of a characterization over another with reference to the *correctness simpliciter* of that characterization. A correctness simpliciter justification is one that does not make clear what claim to usefulness the chosen conceptualization has.

### *5.3.2 Engineering in practice*

What does conceptual engineering look like in practice? As an example, consider the medical research paper "Restoring Study 329 […]" (Le Noury et al. 2015) that re-investigated data from a famous anti-depressant trial, the results of which were first reported in publication in 2001 (Keller et al. 2001). The original publication by Keller et al. concluded that the anti-depressant paroxetine is safe and effective in adolescents. According to Le Noury et al., however, their reanalysis showed no advantage of paroxetine over placebo and indicated significantly higher harmfulness of paroxetine than what was reported in the original journal publication. How can the *same data* support such inconsistent conclusions? Part of the reason for the divergent conclusions regarding harmfulness is that the conceptualization of harms in the reanalysis is different from that in the original publication. The conceptualizations are different in the sense that the two analyses identify different subclasses of harms and use different rules to assign events to the extension of these classes.

For example, the reanalysis classified suicidal events as "suicidal ideation" or "self-harm/attempted suicide" while the original publication denoted such events to the broader and more neutral class "emotional lability".[106] Clearly, the former way of denoting suicidal events has potential to lead to more alarming conclusions than the latter.

Importantly for present purposes, the authors of the reanalysis clearly justify their choice of conceptualization of harms. According to Le Noury et al., their chosen system aligns with the most commonly used system in their field, and it corresponds to clinician descriptions of events in the trial better than the system used in the original publication. In addition, the classificatory system of the original publication is no longer used in their scientific field. Recasting such reasoning in conceptual engineering terms, Le Noury et al. argue for the appropriateness of their chosen conceptualization in terms of explicitly stated criteria: comparability-supporting congruence with widely endorsed practice, and faithfulness to conceptualizations used throughout the trial. Because the conceptual system and the decision criteria are explicitly argued for, it is easy to review and debate their appropriateness, something that the authors of the reanalysis constantly invite readers and other analysts to do.

That last point, the invitation for others to explore the data in light of other categorizations of harms, is worth highlighting. Although Le Noury et al. argue that their presentation of adverse events is in some respects better than that of the original publication, there is no pretension that the conceptualization Le Noury et al. selected is uniquely appropriate. The same set of adverse events can be divided up and assigned to the extension of types of harms in a plurality of ways, each of which may be indexed to a valuable purpose. Le Noury et al. show tolerance towards a variety of conceptual systems and criteria for choosing among them.

As this example shows, the revision and creation of concepts need not be revolutionary in order to count as conceptual engineering. In the case of "Restoring Study 329", the new conceptual system amounts to a partly new conceptualization of subtypes of harms and a novel way of assigning certain adverse events to the extension of those subtypes, *relative to the original study.* For example, Le Noury et al. present self-harm related adverse

[106] See especially sections "Coding of adverse events"," Coding and representation of adverse event data", "Principal findings and comparison with original journal publication" and Table 5 in Le Noury et al. (2015).

events in a new way. In concept-speak, Le Noury et al. place self-harm related adverse events in the extension of a concept that does not have the same intension as the concept that denoted the same adverse events in the original study. Such changes are subtle but consequential, as can be seen from the fact that the two studies arrive at very different conclusions about the harmfulness of paroxetine.

Such changes, although consequential, may seem too small to count as the creation of new concepts. But historical and psychological work on conceptual change suggests that *even in cases where conceptual innovations seem like revolutions*, the actual process that leads to the creation of new concepts is best characterized as an extended, iterative series of incremental changes (Nersessian 2008). The change appears revolutionary only when one compares two concepts that are many steps apart from each other, but the non-revolutionary steps are what constitute the revolution. If conceptual engineering were about conjuring up new concepts from thin air, it would not be a very practicable outlook on concept creation.

The impetus to revising a concept (or a system of concepts) can come from any number of sources. Arguably, two of the most common motivations are: i) new empirical knowledge of the extension of the concept, and ii) intuitions (broadly construed) about the appropriate usage of the concept. These two motivations to revision are discussed in sections 5.4 and 5.5, respectively.

## 5.4 Conceptual engineers carving nature's joints

### *5.4.1 The worry*

To many philosophers, Carnap and conceptual engineering connote construction, and construction, in turn, entails philosophical trouble. When Carnap and the conceptual engineers go around constructing concepts, the sceptic thinks, they forget the same thing most construction-minded philosophers seem to forget: reality.[107] Shouldn't our conceptual apparatuses mirror our knowledge of nature and reality, not the whims of self-

[107] This reading of Carnap's approach to concept construction is present in e.g. Laurence and Margolis (2003, 254). They state that in early 20th century "it was widely thought, following Carnap and others, that scientific concepts must be definable a priori" (ibid.). I don't have the space here to elaborate on my reading of Carnap, but I think much of his writing emphasizes that empirical inquiry informs concept formation (see. e.g. Carnap 1950b). Already in his (1934, 19), Carnap wrote that: "what induces us to prefer certain language form to others is the recourse to the empirical material which scientific investigation furnishes".

proclaimed conceptual experts? Shouldn't we aim to carve nature at its joints and capture genuine natural kinds, not proliferate artificial and arbitrary constructions? Aren't good concepts discovered rather than constructed?

As I indicated in the chapter introduction, the issue of construction is hairy and discussed in a multitude of philosophical sub-literatures. Instead of diving deep into the theories, I propose we look at *examples* of how conceptual engineering navigates "push-back from reality", that is, how the discovery and (usually empirical) study of entities, attributes and laws influences the engineering of concepts.[108]

Speaking in theoretical terms, the following examples illustrate the fact that while the concept (the idea) is constructed, the attribute or entity in the extension of the concept[109] frequently has some kind of independent existence relative to the concept-holder(s). Independent existence here means, simply, that the nature of entities or attributes is not fully in control of the concept-holder. Put in yet another way, the study of the concept's (alleged) extension can surprise the concept-holder, because she does not control every aspect of the relevant entities or attributes.[110] The push-back from reality emerges when we hold erroneous (or otherwise unproductive) beliefs about the nature of entities and attributes that have such independent existence, that is, when the extension (or its vicinity)[111] surprises us. The following examples make this general characterization concrete.

### *5.4.2 Reality pushes back I: the case of the anomalous entity*

The platypus is a quiet-loving animal that lives in Australian freshwater lakes and streams. When European settlers first encountered platypus in the late 18th century, it baffled

[108] In an important sense, of course, social pressures also exert *real* pressure on concept formation. For example, if people refuse to use the concept of preferences to denote choices (as some economists would have it), that is very real push-back to a conceptual apparatus. Although real and important, I will exclude such push-back from the purview of the present analysis, because it pertains to the usage of the concept, not to the concept itself, which is of interest in this section.

[109] Or close to the extension of the concept, as in case I below.

[110] Such a conceptualization of "independent existence" applies to entities and attributes that (likely) would exist whether or not the human race existed (like the entity octopus or the attribute mass) as well as many entities and attributes that humans have somehow shaped or brought to existence (say, the entities poodle and AlphaGo or the attribute depressed). This conception of independent existence leaves out entities and attributes that are in every aspect of their existence dependent on human beings, say, the mental image of Sisyphus I have while writing this. I think my notion of independent existence has similar connotations as Chang's notion of reality (see his 2017b).

[111] See again the first example below.

them: the creature had fur like a mammal, but it laid eggs and had webbed feet and a bill, making it seem more bird-like or reptile-like than mammal-like. Many European scientists thought the descriptions and samples of platypuses that arrived from Australia manifested a hoax rather than a new discovery – so odd was the animal.

Of all its odd features, what interests us is the following combination: platypuses have mammary glands and lay eggs. Although in current classificatory practice this combination is not anomalous (platypuses are classified as mammals), it certainly was at the time of the discovery of the platypus (Moyal 2001). Hall (1999, 212) writes:

> *In the taxonomy established for European species by European naturalists, it was axiomatic that all milk-producing animals give birth to live young, and so, by definition, are mammals. Warm-blooded egg-laying animals were birds. Cold-blooded egg-laying animals were reptiles. There was no place in this scheme for the platypus.*

The discovery of platypus challenged established classificatory practice by showing that there was no place for a "piece of reality", i.e. the platypus, in the scientists' conceptual apparatus. The intension and extension of the scientists' (implicit) concept of mammal had to be rethought in view of the discovery of the anomalous animal. Ignoring the historical subtleties for the sake of the argument, we can say that the extension of "mammal BP" (before platypus) included only milk-producing animals that give birth to live young, while the extension of "mammal AP" (after platypus) included milk-producing animals whether or not they give birth to live young.

The move from mammal BP to mammal AP was certainly motivated by push-back from reality, but it was not *determined* by it. There are infinitely many alternative ways to re-engineer the classificatory system to accommodate the platypus, and a number of them were in fact proposed. For example, it was suggested that the platypus might belong to a new non-mammalian class or a new intermediate class between reptiles and mammals. In other words, although reality (or nature) did set constraints, it did not force the conceptual engineer's (i.e. the scientist's) hand. The case of anomalous entity is not a counterargument

to conceptual engineering, but an example of the constraints conceptual engineers have to navigate.

### *5.4.3 Reality pushes back II: the case of empty extension*

Many philosophers think that diseases have to have a mechanistic, biological foundation in order to count as genuine diseases (e.g. Stegenga 2018, sec. 2.4). In other words, a cluster of harmful symptoms only belongs in the extension of the concept of disease if those symptoms are produced by a biological mechanism. On this account, for example, depression is a disease only if the observable symptoms associated with it (depressive mood, anxiety, insomnia and so forth) are produced by some common mechanism, for example, a particular dysfunction in a particular neurotransmitter system.

Suppose we buy this conception of disease and set out to study depression as such a biologically grounded disease. Most depression researchers today believe that the observed symptoms of depression are not caused by a single shared biological mechanism, but rather by a myriad of different and complex mechanisms that combine environmental, biological, phenomenological and all sorts of other factors (e.g. Beck and Alford 2009). If this is true, scientists will (continue to) fail to discover anything that corresponds to the concept of depression *qua* set of symptoms caused by a shared biological mechanism. Put in other words, the extension of *this* concept of depression is empty, i.e. nothing instantiates the conditions it sets as requirements for something to count as depression.

Does this mean that reality (or our knowledge of reality) forces us to abandon the concept of depression – has reality told us there is no depression? Of course not. Although reality has genuinely taught us something about the extension of one concept of depression (that it is empty!), reality does not force a conceptual engineer's hand. She has plenty of avenues open, for example: re-clustering the symptoms of depression into smaller subgroups and isolating biological mechanisms driving each type of depression; conceptualizing depression in terms of observed symptoms and accepting that it is not a disease (but nonetheless a useful category); or conceptualizing depression in terms of observed symptoms and changing the concept of disease so that diseases do not need to have a biological foundation. Reality has nothing to say about the *correctness* of any of these solutions (although reality does bear on the *usefulness* of each solution). It is up to the maker and user

of the conceptual apparatus to accommodate the fact that a particular conception of depression turned out to have an empty extension. Reality sets constraints but does not do the conceptual engineer's job.

### *5.4.4 Reality pushes back III: the case of heterogeneous extension*

Factor analysis, the psychometric method we described in chapter 2, can be recast as a method of discovering heterogeneity in the extension of a concept. Consider the following example. A psychometrician starts off by formulating the target concept, say well-being, and theorizes that it denotes a homogeneous attribute that changes in specific, regular ways in response to life events (say, divorce diminishes well-being, having children increases it, and so forth). She formulates a battery of questions that, to the best of her (and her colleagues') knowledge, track that target concept. The test is administered, and the data factor analysed. It turns out that the items load rather neatly on two factors, that is, there are two distinct aspects that "explain" the correlations the test items have with each other. An interpretation of the factor analytic results is proposed: the test tracks two aspects of well-being, affective and cognitive. In other words, some questions tap into feelings of happiness (i.e. affective aspect of well-being), while other questions set respondents in a more analytic, evaluative mode vis-à-vis their well-being (cognitive aspect of well-being). Upon further investigation, it turns out that the affective and the cognitive aspect obey different psychological regularities, that is, they change in distinct ways in response to life events. Say, for example, having children increases the cognitive aspect of well-being, but diminishes the affective side.[112]

From the perspective of concept formation, such a factor-analytic result indicates that, what might have been thought of as a concept that denotes a simple, unitary and homogeneous attribute (well-being) turns out to encompass fairly distinct sub-attributes (affective and cognitive aspects) that obey different laws or tendencies. In other words, reality pushes back on the idea that the extension of our concept of well-being is homogeneous and obeys simple regularities. The conceptual engineer must re-engineering her concept of well-being to hold on to the idea of homogeneous extension, or revise the extension of her concept

[112] The history of how type I and type II diabetes got distinguished from each other is similar to the hypothetical case I outlined here. Chang makes a similar point using diabetes as an example, see his (2017a).

in heterogeneous terms. Reality pushes back, and a good conceptual engineer makes her moves mindful of the push-back.

### *5.4.5 No mindless construction*

These examples illustrate that empirical and statistical study of entities and attributes can push us to revise concepts in a multitude of ways.[113] Sometimes a new entity surprises scientists, because it does not fit neatly into any of the conceptual drawers that suggest themselves as potentially applicable (e.g. platypus not fitting into classificatory apparatus). At other times the extension of a concept surprises scientists by being empty, that is, nothing instantiates the concept (e.g. when depression *qua* biologically grounded disease cannot be empirically discovered). And sometimes the extension is unexpectedly heterogeneous (e.g. when well-being turns out to have multiple aspects to it). The examples could be multiplied, but the gist ought to be clear: when scientists are surprised by attributes and entities that are (or were thought to be) in the extension of their conceptual apparatus, this frequently leads them to revise their conceptual apparatuses.

I think this should assure the reader that conceptual engineering is not mindless construction of artificial and arbitrary conceptual systems. That said, it is perhaps useful to remind that the push-back from reality often comes from research on attributes and entities that have arguably been shaped or partially constructed by human beings. Although the attribute anorexic and the entities poodle and AlphaGo have been brought to existence via human activity, they can nonetheless be in the extension of useful concepts, and new knowledge about them can push us to revise our conceptual system.[114] Put in yet another way, useful concepts need not denote nature's joints, if nature's joints are meant to be perfectly human-independent. That useful concepts need not denote nature's joints in this

---

[113] Perhaps the most common way in which reality pushes back on the efforts of conceptual engineers is what we might call the case of mistaken attribution. Mistaken attribution occurs when we have a well-defined, well-justified concept and we mistakenly identify an entity (or an attribute of an entity) as belonging to the extension of that concept. When a doctor misdiagnoses a patient, a biologist misidentifies a milk-producing animal as a reptile, a (bad) cook chops zucchini trying to make a cucumber salad – in each of these cases, an entity (or a feature of an entity) is wrongly placed in the extension of a concept it does not belong to. Such instances usually do not lead to the revision of the concept if the concept is well-established and useful. It is simpler to just relocate the misplaced item in the established conceptual drawer it belongs to, than to rip open the conceptual apparatus that is supported by convention and continuity and whatever other valuable aims the existing conceptual organization exhibits.

[114] See the notion of "independent existence" described above – it clearly leaves room for such construction.

sense is easy to see from the (earlier-mentioned) fact that useful concepts sometimes denote nothing at all. Many concepts have an empty denotation, particularly in logic and mathematics, but equally in any science that deals with so-called ideal types, such as perfectly rational human-beings or frictionless planes. Because such concepts do not capture anything, they cannot capture nature's joints. But their ubiquity testifies to their usefulness.

## 5.5 Conceptual engineers pumping intuitions

### *5.5.1 Intuitions are not privileged*

We started off noting that there are two dominant debates in the philosophical literature on concept formation: one regarding construction and one regarding intuitions. We have now seen how conceptual engineering navigates construction and push-back from reality. In this section we will turn to intuitions, and the central place they have in the (arguably) most popular philosophical approach to concept formation: conceptual analysis. I shall argue that conceptual engineering carves a middle path between conceptual analysis and its severest critics.

Conceptual analysis comes in many stripes (see e.g. Grice 1989; Jackson 1998), but Kitcher (2012, 196) provides a helpful summary:

> *You start with an everyday concept […] and a bold innovator proposes an analysis of that concept, laying down conditions that are intended to be necessary and sufficient. Others react to the proposal by questioning the terms used in providing the analysis (urging that they are unclear, inexact, or whatever) and, most popularly, by putting forward examples intended to show that the suggested conditions are not necessary or not sufficient. These examples are sometimes grounded in ordinary usage about more-or-less ordinary situations, sometimes in predictions about ordinary usage with respect to quite extraordinary (even bizarre) situations. They may force a long series of revisions to the analysis originally proposed.*

The term "intuition" does not appear in this description, but it is implied by talk of "ordinary usage" of a concept in various ordinary and extraordinary situations.[115] The situation is set forth in the thought experiment (say, the setting of Nozick's experiment machine or the Gettierian setting of real and fake barns) and we use our intuitions to decide whether or not the target concept applies in that situation (i.e. whether the person plugged into the experience machine is *well* or whether the person identifying a real barn among fake barns has *knowledge*). On this account, intuitions are not metaphysically mysterious or epistemically complex but almost the opposite: broadly speaking, to have an intuition about the usage of a concept is to regard its applicability in a given situation as *prima facie* acceptable – intuitive, in a word (I follow (Chalmers 2014) with such a minimal account of intuitions).[116] To pump one's conceptual intuitions, then, means consulting one's internal faculties, say thoughts and feelings, about one's willingness to apply a concept to a specific setting.

Conceptual analysis has been criticized from various angles. For example, it has been argued that a concept need not be structured in terms of necessary and sufficient conditions but can rather be characterized in terms of family resemblances, prototypes or with reference to the concept's place in a theory (Wittgenstein 2009 [1953]; Putnam 1970; Kuhn 1974). By contrast, the burgeoning field of experimental philosophy has revealed that philosophers' armchair intuitions about intensions and extensions are not nearly as widely (let alone universally) shared as conceptual analysts have tended to think (Weinberg, Nichols, and Stich 2001; Machery et al. 2004). Moreover, the grounds are arguably thin for claiming that

---

[115] Philosophers disagree on what intuitions are, and consequently, on whether there are significant mismatches between conceptual intuitions and ordinary usage of concepts (see Cappelen (2012) and discussion in *Philosophical Studies* 2014, 171(3)). Under some interpretations, conceptual intuitions coincide with conceptual usage in most of the interesting cases (Eklund 2015). In what follows I shall discuss ordinary usage and conceptual intuitions simultaneously. This seems to be a common move in recent Carnap scholarship: for example, Kitcher (2012) and Shepherd and Justus (2015) speak about everyday usages and conceptual intuitions simultaneously in their accounts of Carnapian explication. Furthermore, I think the following argument applies with minor modifications even if there were significant mismatches between conceptual intuitions and everyday usages. The proposed reasoning is, in other words, robust across many (but not all) specifications of the relationship between concept usage and conceptual intuitions.

[116] The word "applicable" is important here. The question I am concerned with is whether or not it is intuitive to apply a concept in a certain (thought experimental) context, not whether the content of the concept is itself intuitive. To see how these two come apart, consider the concept of knowledge. Many philosophers find it intuitive that the concept of knowledge does not apply in the Gettier cases. But when we revise the traditional definition of knowledge ("justified true belief"), the definition becomes more technical, convoluted, harder to understand and apply – less intuitive, one might be tempted to say. This chapter deals with the usefulness of intuitions about applicability, not the content of the concepts.

philosophers' conceptual intuitions are privileged, epistemically or otherwise (Weinberg et al. 2010).

More fundamentally, some critics are asking whether intuition pumping (via thought experiments) serves any valuable function in establishing a concept's appropriate intension and extension. Although even the severest critiques of intuition-based philosophical methods agree that intuitions can be a source of information about how people actually employ concepts (cf. Hintikka 1999), many contemporary philosophers are now asking: what is so great about knowing how people employ concepts? Why should we think that the correct or best characterization of a concept is one that coheres with folk or philosophers' intuitions, especially since folk tend to have divergent and contradicting intuitions about conceptual content? (Weinberg, Nichols and Stitch 2001; Kitcher 2012; Scharp 2013; Eklund 2015; Shepherd and Justus 2015)

Although these objections are well rehearsed in the literature, intuition consulting is still philosophers' main tool of arriving at a characterization of their target concept. Intuitions are being pumped here, there and everywhere in philosophical literatures that try to determine a characterization of concepts such as disease (Cooper 2002; Kingma 2010; Stegenga 2018), well-being (Rice 2013; S. M. Campbell 2016; Hausman 2015) and others. We are thus left with an awkward predicament: methodologists act as if they are successful at eroding the justification of conceptual analysis, while philosophers working on particular concepts act as if they are successful at forming concepts by means of conceptual analysis.

Conceptual engineering carves something of a middle path between conceptual analysts and their (severest) critics. Under the conceptual engineering approach, the results of the ubiquitous intuition pumps have no *privileged* claim to acceptability (that is, no more claim than any other conceptualization) *unless a plausible argument can be constructed to the effect that, in that context, an appropriate and useful concept corresponds to philosophical intuitions.* For example, say one is contemplating whether to define well-being in terms of pleasurable experiences, and a Nozickian attacks with the experience machine thought experiment. Or a Gettierian corners one with a thought-experimental set of barns to question one's definition of knowledge. A good conceptual engineer answers questions of expediency before conceding to either intuition pump. Is congruence with the pumped intuitions useful

in this context? Will the connection between pleasure and well-being break down in the context in which the concept will be used? Is "true justified belief" a workable characterization, even though it fails to be intuitive in every imaginable situation? Should some intuitiveness be traded off to increase some other aim, say, simplicity? Is maximal intuitiveness less important than, say, measurability? And so on for other similar questions of expediency.

Conceptual engineering thus navigates the tension regarding traditional philosophical methods by granting conceptual analysts that intuitions can be useful, but by simultaneously insisting that the usefulness of conceptual intuitions needs to be balanced and argued for vis-à-vis other aims and criteria. By re-characterizing concept formation in terms of trade-offs between context-specific valuable aims, conceptual engineering grants the *critics* of conceptual analysis that intuitions are not the end-all of concept formation, and that indeed often the most usable concept is one that sacrifices intuitiveness for some other valuable aim, say, exactness.

A conceptual analyst might retort that conceptual engineers are just reinventing the wheels conceptual analysts have been using all along. Intuitions are being pumped, so the argument goes, precisely because that produces maximally useful concepts! For that claim to be credible, though, one would have to argue that intuitiveness is conducive to usefulness in the various conceptual fields where conceptual analysis is and has been applied. Many of our conceptual needs are such that everyday intuitions are *prima facie* useless – think of mathematical concepts and concepts needed for the purposes of quantum physics, or concepts relating to AI or high arts. Of course, mathematicians and physicists might have claim to useful intuitions in their respective subfields of mathematics and physics, due to their (tacit) knowledge and skills in the field. [117] Even so, not all intuitions that enter a physicist's head count as advances in quantum physics, and it is rarely, if ever, the case that intuitiveness alone makes a scientific concept successful. Rather, as the growing literature on the role of values in scientific discovery testifies, triumphant conceptual apparatuses exhibit various features that are deemed useful, such as completeness, measurability, mathematical tractability and so on. Conceptual engineering comes apart from conceptual analysis in recognizing that there

[117] The purview of philosophical expertise is much more difficult to outline, and therefore appeal to expert intuitions is much harder to justify in philosophical contexts (see Weinberg et al. 2010 for a discussion).

is much more to useful concepts than intuitiveness. Hence if conceptual engineering gains currency in philosophy, intuition pumping will likely decrease dramatically.

### *5.5.2 Intuitions are still welcome*

All that said, one should not rush to the other extreme and claim that intuitions have no place in the formation of useful concepts. In recent methodological literature, the difference between conceptual engineering and conceptual analysis has sometimes been presented in a manner that suggests such an extreme approach, particularly with respect to folk intuitions (i.e. the intuitions of laypeople as opposed to people with a specific expertise, say, philosophy or biology). Crudely put, the idea is that conceptual analysis is all about (folk) intuitions, while conceptual engineering is all or mainly about ignoring (folk) intuitions for the sake of usefulness. Implicit in these arguments is the idea that intuitive concepts are rarely the most useful ones, especially for scientific purposes. Eklund (2015, 378), for example, writes that: "However we best go about the project of conceptual engineering, it is hardly via relying on competence intuitions[118], or by studying the judgments of the folk." In my view, a similar disvaluing of the usefulness of folk intuitions and laypeople's concept usage is present in e.g. Kitcher (2012, ch. 8), Shepherd and Justus (2015) and Duthil Novaes and Reck (2017).

I think these claims are a bit too severe, because it does seem that intuitiveness can be useful at least *sometimes*. I have already used the example that continuity with conceptual intuitions serves clarity, and it also saves time and effort in laying out a characterization. In addition, and most significantly, I think there are many reasons why valuative or value-laden concepts, such as well-being, progress and development, should preserve people's everyday intuitions, in order to be useful for policy and other normative purposes. The rest of this section outlines these reasons.

There are, firstly, epistemological reasons for believing that valuative concepts should often resonate well with folk intuitions. Rudolf Carnap recognized this, arguing that it is useful to study everyday usage of a concept, because that usage tends to suggest efficient and purposeful ways of conceptualizing things (Carnap 1955). Presumably the rationale for

[118] "Competence intuition" is Eklund's own technical term and denotes (very roughly) intuition that is reliable due to being tied to certain shared linguistic practices in a community of speakers (where community of speakers might be people on Earth or people on Twin Earth).

this is that concepts would not persist in everyday communication, whether in science or outside of it, if those concepts were useless or inefficient, and thus everyday usage provides a kind of short cut to useful concepts. This epistemological benefit of everyday usage applies to some value-laden concepts particularly well: people have privileged understanding of their own lives and values, [119] and arguably the way they conceptualize things, the way they use and apply concepts, to some extent reflects such knowledge.[120] If a normative concept is supposed to serve the formation of policies and normative guidelines that aim for the betterment of the lives of people (in some respect), it makes sense to tap into the privileged epistemic access that people have to their own lives.

The epistemic benefit of tapping into laypeople's knowledge of their own lives is a core motivation for so-called participatory action research in the social sciences. Participatory action research employs a variety of tools and methods, but it always engages and encourages collaboration with members of a target community in order to solve the questions and problems that are significant for that community (see Reason and Bradbury 2008). Consider, as an example, the World Bank's report *Voices of the Poor – Crying out for Change*, which begins with a remark that is telling of the report's participatory approach: "There are 2.8 billion poverty experts, the poor themselves" (Narayan et al. 2000, 2). One of the methods Narayan et al. used to tap into this expertise was to ask the poor to define and reflect upon the meanings they attach to value-laden notions such as well-being and ill-being.[121] The researchers used the emerging definitions and characterizations to outline policy implications, building on the notions of well-being that the poor themselves used and considered significant. In building normative concepts so that they mirror people's personal notions of well-being (and other such concepts), participatory research implicitly privileges the everyday judgments that people make about these value-laden concepts.

One might suspect that the epistemological justification applies to a very limited number of concepts, and that the above argument is therefore not particularly consequential.

---

[119] Note that such knowledge need not concern the truth or falsity of the value judgments that people make to express their value commitments. It can be knowledge concerning one's commitment to certain kinds of values that reflect certain aspects of one's life, and knowledge that these values are appropriate for one's life. Thus, the kind of epistemic privilege we are dealing with does not commit one to a cognitivist stance on value statements.
[120] Haybron and Tiberius (2015) argue that people tend to have a good grasp of appropriate values that pertain to their welfare.
[121] In what follows I focus on the concept of well-being, but similar considerations arguably apply to the concept of happiness. See Haybron (2003).

It is nonetheless reasonable to think that the case of well-being warrants some generalizability to the epistemological argument, because many normative concepts "bottom out" in well-being (cf. Alexandrova 2016). In other words, many normative concepts, such as progress, development and growth (and potentially, health) are partly defined in terms of or in relation to well-being. Thus, the emphasis on everyday intuitions in the formation of the concept of well-being transfers to many other normative concepts. But the epistemological argument has some limitations: it is not always the case that everyday usages are epistemically privileged even in the case of well-being. Sometimes subjects are not epistemically privileged, for example if subjects are misinformed.

A stronger and more widely applicable motivation for intuitive concepts has a grounding in ethics. When we are dealing with value-laden concepts that play a role in normative guidance, fixing values without consulting people who are affected by the resulting normative guidance amounts to an imposition of values (Haybron and Alexandrova 2013; Alexandrova 2016). Such imposition is objectionable on democratic grounds. To give an example of the ethical argument, consider a recent but famous paper by Benjamin, Heffetz, Kimball and Szembrot (2014). They set out to construct a well-being index that combines different aspects of well-being on the basis of people's preferences over these aspects. To do this, they come up with a survey that asks people to make trade-offs between aspects such "happiness" "not feeling anxious" and "sense of achievement and excellence", which are all possible components of the concept of well-being. The exact details of this procedure do not matter here, but the point is that Benjamin et al. (2014) build their scientific concept of well-being on the judgments that people make about things they value. They think that this is an attractive procedure, because it is not paternalistic about what well-being amounts to. In other words, Benjamin et al. rely on an ethical guideline that leads them to prioritize folk intuitions.[122]

It is noteworthy that Benjamin et al. respect laypeople's usage of the concept of well-being while simultaneously pursuing a goal that is typically thought to require a clearly

[122] Such respect for laypeople's notions of well-being occurs in other branches of well-being science as well. For example, some researchers studying subjective well-being argue that their concept of well-being, and the measures that capture that concept (i.e. questionnaires asking people to express their agreement on a rating scale with respect to questions such as "Is your life close to your own ideal?") allow people to "define well-being for themselves" (Diener, Sapyta, and Suh 1998, 35; Helliwell, Layard, and Sachs 2017, 123).

delineated target concept, namely, measurement. Their approach gives reason to believe that it is not only desirable to respect laypeople's notions of well-being, but that it is also possible to do serious scientific work with concepts that closely mirror people's conceptual intuitions. In conclusion, normative concepts illustrate that intuitiveness is sometimes a useful feature of a conceptual apparatus. Hence, we should not rush to shut down all intuition-pumping.

### 5.6 Anything goes?

In this chapter, I have defined conceptual engineering and defended it against several objections and worries. In particular, I have shown that conceptual engineering is not blind constructivism, but rather an approach that is genuinely responsive to discoveries about nature and reality. I also described how conceptual engineering relates to and improves upon conceptual analysis.

Before we put conceptual engineering to work on psychometrics, one final objection needs to be heard. One might object that conceptual engineering leaves us with nothing to go by in concept formation and is therefore as useless a method as Feyerabend's "anything goes". It is true that going on and on about usefulness is not going to get one far in creating and using concepts. Rather, one must roll up one's sleeves and do stuff: pump intuitions, draw up distinctions, derive consequences, empirically explore extensions and more. The next chapter will show how to do this in psychometrics.

# 6. Engineering psychometrics

## 6.1 Operationalism after all

If psychometricians, as per chapter 4, are not producing quantitative representations of non-operationally defined concepts, what exactly is the business they are in?

The diversity of psychometric techniques described in chapter 2 allows for multiple answers. One answer is that psychometrics is about following the rules of psychometrics, as they are laid out in textbooks, APA guidelines, peer reviewers' comments and so on. What psychometricians do, on this view, is negotiating and setting up standards, and then mutually enforcing them. This answer does not even begin to be satisfactory (to critics nor to psychometricians, I presume). Surely standards must have some motivation, some rationale of existence, that goes beyond other standards? Another answer to the question "What exactly are psychometricians doing, if not quantitative non-operational representation?" is this: they are doing whatever works. On this view, psychometricians mix and match techniques from their shared toolbox to make their instruments work in educational testing, in policy guidance, in hiring contexts and elsewhere. This may be a correct surface level answer, but it leaves a big question unanswered: what is it for psychometric instruments *to work* in these contexts, if they are not representing the target concepts?

Another answer, which has in fact been lurking in the background of much of this dissertation, is this: psychometricians are practicing an operationalist approach to measurement. This answer has made many 20$^{th}$ century psychometricians uneasy, and my impression is that it continues to make present day psychometricians uneasy. It is not a characterization that gets welcomed with open arms, as I will show in a moment. Nonetheless, operationalism keeps coming up in the psychometric literature that has been published over the last 90-or-so years. It is therefore worth revisiting psychometricians' relation to operationalism in this final substantive chapter of the dissertation.

Building on an historical overview, I will defend a normative claim, which I call Validation Dualism.

Validation Dualism: *There are two distinct, defensible approaches to psychometric validation, the respectful operationalist approach and the inferentialist approach*.

Respectful operationalism, as I define it, uses psychometric tools to simultaneously construct a test and engineer an operationally characterized target concept. The result is a test that represents relations between test takers in operational terms – in other words, the resulting psychometric instrument represents relations in terms of test responses. This process is *respectful* when it takes into consideration the associations people have about the purported target concept, be it well-being, depression or something else. Inferentialism, by contrast, uses psychometric tools to engineer a non-operationally characterized target concept. The result is a test that represents relations between test takers in non-operational terms – that is to say, the resulting psychometric test yields representations of what underlies or determines test responses. While both inferentialism and respectful operationalism are useful and defensible, combining them unreflectively leads to epistemic havoc – or so I will argue.

This chapter is organized as follows. Section 6.2 gives a brief historical outline of operationalist ideas in psychometrics and defends a version of operationalism. Section 6.3 explicates and defends two approaches to psychometrics: respectful operationalism and inferentialism. Section 6.4 shows that mixing respectful operationalism and inferentialism leads to epistemic havoc. Section 6.5 concludes.

## 6.2 A methodological underdog

### *6.2.1 A brief history of operationalism*

Nobel Prize winning physicist Percy W. Bridgman is typically credited as the father of operationalism (Chang 2009). His most famous, and most often criticized declaration of operationalism appeared in his (1927, 5) book *The Logic of Modern Physics*:

> *In general, we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations.*

This statement may be read as a radically reductionist and counterintuitive theory of meaning: there is nothing more to the meaning of concepts such as length than the operations used to measure those concepts. In other publications, Bridgman defends a less radical form of operationalism. According to Chang (2017b), Bridgman's less radical operationalism stands for the view that scientists should be careful not to assume that different operations measure the same thing, even when the same name or term is applied to the two operations. In this less extreme form, operationalism is not primarily a theory of meaning but rather a methodological outlook or a research heuristic.

One of the main reasons Bridgman opted for operationalism was his desire to make science "safe". Safety has various connotations, but in the context of operationalism, safety is often taken to refer to intersubjective agreement and repeatability of inferences.[123] In other words, defining concepts in terms of operations is meant to eliminate the kind of error and disagreement that tends to accompany inferences to non-operational concepts. Hull (1968, 438-39) summarizes the desire for safety thus:

> *If a scientific concept is synonymous with a set of operations, and if these operations are such that they can be performed publicly by any qualified person, then the intersubjectivity and repeatability so important to the objectivity of science are guaranteed.*

Some logical empiricists, such as Carl Hempel and Herbert Feigl, were intrigued by Bridgman's operationalism. The operationalist focus on observability and testability resonated with logical empiricists' own efforts of putting the language of science on a firm, verifiable or confirmable basis. As is widely known, Carnap dreamed of a language of science, in which all statements are either testable by relatively uncontroversial observational procedures or reducible to statements that are testable in such a manner. However, operationalism differed from logical empiricism with respect to the unit of analysis: where operationalism focused on the meaning of *concepts*, logical empiricism typically looked at the

[123] Bridgman himself thought that operations are a matter of private experience (Chang 2009, 2.4). In doing so he departed radically from other operationalists, who thought of operations as something public and therefore as a grounding of intersubjective agreement. I shall focus on this latter conception of operations and operationalism, because it is more pertinent to psychometrics.

empirical meaning of *statements* (Hempel 1954). Nonetheless, logical empiricism had many similarities with operationalism (at least on the surface), and proponents of the two approaches exchanged ideas enthusiastically.

While Bridgman is often described as the father of operationalism, and logical empiricists are credited for importing the approach to philosophy, in psychology the seeds of operationalist thinking had been planted before either of these approaches emerged. According to Hull, operationalist ideas surfaced in psychology some ten years before Bridgman's famous treatment of physics – but the term psychologists used for these ideas was *behaviourism* rather than operationalism. In his 1913 article, John B. Watson introduced behaviourism in reaction to Freudian psychology and other then-fashionable psychological schools, which focused on consciousness and other unobservable attributes. Watson suggested discarding reference to consciousness and related unobservables, because there is no agreement on what those concepts stand for and how to study them:

> *The time seems to have come when psychology must discard all reference to consciousness; when it need no longer delude itself into thinking that it is making mental states the object of observation.* (Watson 1913, 249)

Instead, according to Watson, psychologists should focus on observable changes in people's behaviour. In particular, Watson's behaviourist psychology consists of reports of reactions in experimental settings: how people respond to visual stimulus or words in a memory test, for example. The operationalist flavour is clear here, that is, memory, vision and other psychological capacities are studied and described in terms of behaviour in a test set-up, a type of operation. [124]

Watson's motivation for behaviourism was *not* that inner states do not exist – they might – but that inferences to internal states are more controversial than reports of behaviour. Two people can easily agree on the answers a patient gives to a personality test but disagree on what those answers mean in terms of the patient's personality and mental

[124] It is not obvious what exactly operations are, but debates about their exact nature do not matter for our purposes. See Bridgman (1959) for a classification.

health. Focus on behaviour will make psychology more transparent and accessible and thereby more useful to other human endeavours, according to Watson:

> *If psychology would follow the plan I suggest, the educator, the physician, the jurist and the business man could utilize our data in a practical way, as soon as we are able, experimentally, to obtain them.* (ibid., 252)

Evidently, Watson's behaviourism was motivated by the value of inferential safety, just like Bridgman's operationalism was.

Operationalism proper, however, was brought to psychology some twenty years later, in the mid-1930s.[125] Its most influential early proponents were Harvard psychologist Stanley Smith Stevens and Berkley psychologist Edward Chase Tolman. It is likely that both Stevens and Tolman were following an implicitly operationalist[126] methodology in their experimental work well before they explicitly declared themselves operationalist (Feest 2005). Tolman, for example, studied the behaviour of rats, in particular, the way the rats' ability to navigate through a maze is influenced by how much and how frequently the rats are fed. Tolman described himself as a behaviourist and was well aware of John B. Watson's work on the conditioning of rats and humans. Given that Watson' behaviourism has been thought of as operationalism under a different name, this indeed suggests that Tolman's work was influenced by proto-operationalist ideas before he explicitly started pursuing operationalism in the mid-1930s.

With their work already implicitly analysable in operationalist terms, it is not surprising that Tolman and Stevens explicitly endorsed operationalism once they encountered Bridgman's and the logical empiricists' writings on the subject. In 1937, Tolman provided an operational analysis of desires (what he called "demands") in the philosophical journal *Erkenntnis* (Tolman 1936). In the article, Tolman suggested that desires can be defined in terms of observed behaviour in a particular test set-up. For example, a rat's desire for food can be defined in terms of how long the rat persists in trying to obtain food in a particular type

[125] Feest (2005) notes that Edwin G. Boring might be credited for bringing operationalism to psychology even earlier than this, in 1920s.
[126] This is frequently called "operationism" in psychology.

of maze. The definition of desire is therefore tied, not just to observed behaviour, but also to the particular test set-up in which that behaviour occurs, in other words, to a kind of operation.

Stevens, who worked within the psychophysical tradition studying subjective auditory experience (among other things), also cited Carnap and Bridgman as the inspiration to his methodological papers on operationalism. Stevens studied how different properties of a tone depend on each other, for example, how the subjective experience of volume of a tone depends on the intensity and frequency of the tone. In one of Stevens' experimental set-ups, the subjects would hear two tones of different frequencies and adjust the intensity of one of the tones until the subject experienced the volumes of the two tones as equal (Stevens 1934). Stevens could then define subjective experience of volume in terms of this operation, that is, the subjective experience of volume could be defined in terms of the judgments of difference and equality, which subjects make when placed in the experimental set-up just described (cf. Stevens 1935). Stevens' motivation for such an operationalist approach was akin to those of Bridgman and Watson: to render psychology and other sciences public and repeatable, and thereby objective (1935, 328).

The work of Stevens and Tolman started discussions on the place of operationalism in psychological methodology. The journal *Psychological Review* offered a platform for many a paper on operationalism in the late 1930s and the early 1940s, and these discussions culminated in a symposium organized in 1945 and sponsored by *Psychological Review*. Contributors to the symposium included behaviourist psychologists (e.g. B. F. Skinner), logical empiricists (e.g. Herbert Feigl) and Bridgman himself, thereby bringing together the three main strands of operationalist thinking. The contributors appear to have had lots of disagreement over what operationalism is and what its role should be in psychology and science more broadly (Green 1992). Despite failure to agree on fundamentals, interest in operationalism persisted for another decade or so. In 1953, another operationalism symposium was organized at the meeting of the *American Association for the Advancement of Science*. By 1959, however, many previously operationally-minded psychologists, such as Tolman, had come to doubt the adequacy of operationalism (Green 1992). Philosophers rejected operationalism together with logical empiricism, viewing both as extreme and

inadequate (Green 1992; Chang 2009).[127] In psychology too, operationalism gradually faded into background in methodological debates.

The contemporary status of operationalism in psychology is best described as a mix of acceptance, abhorrence and confusion. The mix of opposing attitudes, acceptance and abhorrence, can be observed, first of all, from the occasional debates that emerged since the fade out of the initial interest in operationalism. These discussions occurred at least in the early 1980s on the pages of the *Journal of Mind and Behavior*, in the early 1990s on the pages of the journal *Theory & Psychology* and again in the beginning of 2000 *in Theory & Psychology* (Feest 2005). Each time there were vehement debates, with some contributors defending operationalism and others objecting with equal force.

It appears that stakeholders not only disagree on whether operationalism is a good thing, but they also disagree on whether psychometrics tends to be operationalist. In the secondary literature, which describes and comments on psychometrics, operationalism is frequently thought to permeate psychometric practice (e.g. Green 1992; Bickhard 2001; Michell 2008). However, in the primary literature psychometricians frequently argue that their measures are intended to capture unobservable, latent attributes, the nature of which is inferred from observed patterns of responses on a test (see section 4.2 above). The latter attitude points to non-operationalism, because tests are used to make inferences to the real attribute of interest, instead of defining the attribute in terms of the operation. I regard it as another example of confusion that some of the classics in psychometric literature have been described both as defences of operationalism and as guidelines for moving away from operationalism. For example, Green (1992) states that Cronbach and Meehl's classic 1955 article is an "attempt to salvage operationism" although the usual reading (which I outlined in section 4.2) is that the article proposes a non-operational methodology for psychometrics.

[127] In philosophy, operationalism was lumped together with logical empiricism and viewed as an account of meaning, which stated roughly that the meaning of a concept is equal to the operations used to measure it. This bundling of the two approaches had the consequence that, once logical empiricists' criteria of meaning was found to be inadequate, operationalism was abandoned as well (Chang 2009). This happened despite the fact that Bridgman and many other operationalists seem not have thought of operationalism as a theory of meaning (at least after due consideration), but rather as a methodology or heuristic for research (Feest 2005, cf. Chang 2009). Either way, philosophers largely abandoned operationalism as a failed philosophy of science.

All this suggests that psychometricians (and commentators) are not single-minded on what the relation between operationalism and contemporary psychology is, and what it should be.

### *6.2.2 Engineering operational concepts*

Operationalism divides opinions because it stands for multiple different things. Even in this condensed historical overview, we have seen operationalism treated as a theory of meaning, as a research heuristic and as a kind of behaviourism (in psychology). There are many more disagreements over the nuances of what operationalism amounts, on questions such as: What are operations? Must all scientific concepts be operationally defined? What are the criteria for distinguishing good and bad operations? (see Green 1992; Chang 2009) This convoluted scene is set for people to talk past each other and to therefore fail to make progress on the topic of operationalism.

Instead of engaging with operationalism directly, I would like to use its intellectually rich history to carve out a defensible, operational approach to conceptual engineering. In other words, I shall outline an approach to engineering concepts, which takes on board insights from defences of operationalism and which therefore has an operationalist flavour. This approach, which I shall simply call *engineering operational concepts*, does not exhaust conceptual engineering, as should be clear based on chapter 5. It is rather a particular application of the conceptual engineering framework.

Operational concepts shall here be understood as concepts that are characterized in terms of an operation, usually a test (e.g. HAM-D) or another kind of experimental set up (e.g. Tolman's rat maze). To be characterized by an operation, a definition of the concept must mention the relevant operation as a crucial requirement for something to count as an instance of that concept. For example, an operational concept of depression could be defined as "a patient is depressed if she has a score higher than 18 on the HAM-D scale". Engineering operational concepts, by extension, amounts to characterizing concepts in terms of tests or experimental set-ups. It is important to keep in mind that we are indeed dealing with definitions here, not truth-apt empirical claims. We should not, for example, read the just-mentioned definition as stating an empirically assessed generalization that "patients

who score higher than 18 of HAM-D tend to be depressed", where depression is defined independent of HAM-D.

To get at a more thorough picture of the engineering of operational concepts, I shall turn to some objections that have traditionally been levelled against operationalist ideas. By showing how the conceptual engineer can respond to these objections, I deepen my characterization of the engineering of operational concepts.

One worry we should discard immediately is that operational analysis is *arbitrary* stipulation. As with all conceptual engineering, the constructed operational concept must be useful, and as discussed in chapter 5, a concept is rarely useful if it does not resonate with the users of the concept. It is typically bad conceptual engineering to go around constructing arbitrary operational definitions, such as "a person is depressed if their body temperature is above 37 degrees Celsius". Most operationally inclined psychologists already accept the requirement for non-arbitrariness. For example, when Frank et al. (1991) set out to provide operational definitions of depression-related concepts such as remission and relapse, the definitions were a result of debate among scientific experts. The engineering of operational concepts need not and should not be arbitrary but well argued for on the basis of considerations over usefulness, as described more thoroughly in chapter 5.

Another common worry is that operationalism is reductive. For example, from some of his remarks it looks as if Bridgman was claiming that the whole meaning of a concept can be reduced to operations – there is nothing more to the meaning of a concept.[128] This is extreme and not what a careful conceptual engineer would say. From the conceptual engineering perspective, it is evident that concepts can be characterized in a variety of ways, and operational characterization is one among many useful characterizations. When defining depression operationally, the conceptual engineer leaves out some non-operational meanings, not because they are irrelevant, but because a single definition cannot pack every connotation a contested concept like depression has. If this is reduction, it is benign, local reduction: the richness of meanings associated with depression is reduced to operational terms for a particular context, for a particular use, for good reasons.

In debates about operationalism, one often hears the charge that operationalism leads to harmful proliferation of concepts (e.g. Hempel 1966; Hull 1968;

[128] Bridgman's later remarks were more nuanced than this (Chang 2009).

Leahey 1980). If every operation defines a new concept, scientists will drown in concepts – or so the objection goes. My reply is two-fold: i) operationalism need not lead to uncontrollable proliferation of concepts, and ii) sometimes having a plurality of operational concepts is a good thing. Firstly, just like in the normal evolution of concepts, new operational concepts can replace old ones. This way there need not be proliferation of concepts but a succession of improving characterizations. Secondly, sometimes we need to have a plurality of operational (and non-operational) concepts that exist side-by-side. This is because meanings associated with contested concepts like depression are not just unruly and many, but sometimes contradictory. [129] For example, there may be good arguments for characterizing depression in terms of a test of serotonin deficiency and good arguments for leaving a test of serotonin levels out altogether. One concept of depression cannot encompass both characterizations, because that would be contradictory. We might therefore opt for a plurality of operational concepts, noting their relations and domains of applicability carefully and transparently.[130]

It is a surprisingly common misconception that operationalism is an anti-realist thesis. It is, in other words, thought that operationalists either deny the test-independent existence of things their measures apparently pertain to or they deny the possibility of epistemic access to test-independent things (Lovett and Hood 2011; Maul and McGrane 2017). On this view, for example, operationalists define depression in terms of a test because they believe either that depression does not exist independent of the test or that there is no way to access depression *qua* test-independent reality.

Is the engineer of operational concepts necessarily an anti-realist? First, the engineer of operational concepts is very much a realist about the entities or attributes the test situation brings about. That is, when people respond to questions on a depression test, those responses, whether communicated non-verbally or verbally, are of course real. They are,

[129] The critic might be tempted to respond that if the meanings attached to depression are contradictory, then some of those meanings are wrong. Put in more general terms, the claim is that we need not proliferate concepts, just point out cases where the concept of depression is used wrong – for example, that people who say that depression is associated with serotonin deficiency are simply wrong. It is, however, difficult to make the case that a concept is wrong without eventually having to fall back on one's own conceptual preferences. That is why conceptual engineering advises us against (what I have previously called) *correctness simpliciter* claims about concepts.

[130] On the other hand, one might argue that it is valuable to find concepts that cohere with each other (Chang 2017b).

however, a kind of test-dependent reality, in the sense that they were brought to fore (or sometimes brought about) by the testing context.

Second, and more interestingly, the engineer of operational concepts can grant that something test-independent[131] often (but not always!) drives or brings about the way people respond to depression test items. That test-independent something might be feelings of hopelessness, serotonin deficiency, a combination of these, or something else. Even if she grants this realist claim, she may want to stick with an operationally defined concept, because she does not (presently) have epistemic access to that "test-independent something", that is, she is not able to make claims about the kind of test-independent reality that drives response behaviour. The conceptual engineer may not know whether that reality is physical or behavioural or mental, whether it is homogeneous or heterogeneous and in what sense, whether it is patterned in a manner that is best represented with quantitative or ordinal or classificatory numeral structures, and so on and so forth. The engineer therefore retreats to claims about responses on a test: "Patient 34 received a high score of depression test T", "Patient 66 received a higher score than Patient 34 on test T", and so on. In such cases, operationalism is motivated by cautiousness, not by anti-realism.[132]

The final worry I wish to consider is the relation between engineering operational concepts and the currently trending mechanistic thinking in the philosophy of social and medical sciences (e.g. Thagard 1999; Machamer, Darden, and Craver 2000; Russo and Williamson 2007; Illari 2011). Focusing on diseases for the sake of brevity, it may seem that operationalists are wedded to the symptoms-based way of conceptualizing diseases. For example, a symptoms-based view of depression would characterize the disease in terms of insomnia, suicide attempts, changes in weight and other similar aspects. Many depression-related test operations, such as self-report questionnaires, capture exactly these kinds of things. The mechanistically minded researcher, however, considers such an approach wrong-headed, because what really matters is what brings about the symptoms, not the symptoms

[131] Test-independence means that a characterization of what the test pertains to need not mention that test. In other words, the test is not a necessary component of an adequate characterization of what is being described with the test. The operationalist always has to say: "The patient is highly depressed in the sense that they receive a high score on test", thereby incorporating her instrument in the characterization of "depression" while the inferentialist makes a more universal claim: "The patient is highly depressed, and this is true whether or not they were tested with test T".

[132] Cautiousness is especially valuable in fields like psychometrics, where test results can have a significant influence on people's lives (as discussed in chapter 1).

themselves. That is, in characterizing diseases, we should capture the mechanism that produces the observable characteristics a test tracks, according to the proponents of mechanistic tinking.[133]

Why would we want mechanistic characterizations rather than operational ones? Consider the following case.[134] Say an operationalist constructs test T that asks questions about depressive mood, anxiety, insomnia and other depression related symptoms. Suppose the operationalist defined depression in terms of this test: depression is what test T measures. Suppose now that in further studies it turns out that two different kinds of mechanisms drive the way people respond on the test. Some people receive high scores because their bodies are unable to produce chemical C that is associated with feelings of happiness and well-being. In other people, the high scores appear to be, not because their bodies cannot produce chemical C, but because their bodies are for some reason unable to utilize chemical C in the normal way (that is, so that it brings about feelings of happiness). The mechanist would say that it is best to re-conceptualize depression so that we differentiate the two sub-types of depression in accordance with the mechanisms that drive observed response patterns: depression that results from failure to produce chemical C (call it "C-less depression") and depression that results from failure to use chemical C (call it "C-block depression").

Why is the new, mechanistic characterization better? One reason is that mechanisms confer a sense of understanding and explanation: it feels like I have a better handle of depression when I know what underlies the surface level responses. Another reason is that mechanisms help us intervene. If it known that a person has a high score on a

---

[133] In what follows I talk about the distinction between observables and unobservables in a manner that might look philosophically unsophisticated. Of course, the dividing line between observables and unobservables is contested. Because the forthcoming discussion is comparative, we can think of the distinction between observables and unobservables in terms of relative epistemic demandingness: claims about unobservables are harder to corroborate and reach intersubjective agreement on than claims about observables.

[134] The case is constructed to mirror the historical development of the concept of diabetes mellitus. Diabetes refers, broadly speaking, to conditions where high blood sugar levels persist in the body for abnormally prolonged periods. Although the overt symptomatology of diabetes has been known for hundreds of years, the causes of the observed characteristics became understood around the beginning of 1900. In particular, it was discovered that similar looking surface problems (blood sugar level not lowering the same way it does for most people) are explained by two different biological mechanisms: in some cases the symptoms are caused by the inability of the pancreas to produce enough insulin while in other cases similar symptoms are caused, not by lack of insulin, but by loss of sensitivity to insulin. This discovery lead to the division of diabetes to type I and type II according to the mechanism that underlies the observed symptomatology.

depression scale because their body fails to produce chemical C, a successful intervention might be one in which the patient is prescribed chemical C. If, however, the patient gets a high score because their body cannot process chemical C, it is probably no use prescribing C. Some other intervention will have to be developed.

There is no denying the value of the mechanistic endeavour. Nonetheless, the mechanistic approach should not be considered a replacement but a complement to operational analysis. First of all, it is all too well known that mechanisms are hard to discover and describe accurately, especially in the social and medical sciences. In the absence of knowledge (or even credible hunches) about underlying mechanisms, operationally characterized concepts may be the only way to go. Secondly, when credible mechanistic reasoning is available, we can harvest the benefits of both approaches by using mechanistic knowledge to formulate better tests, in terms of which the concept is re-characterized. Consider, for example, a concept of depression defined in terms of the following two-stage test: i) administer psychometric test, pick out high-scoring patients and prescribe them chemical C, ii) after an appropriate amount of time, administer psychometric test again for those who were prescribed chemical C, and separate high scoring patients from the low scoring ones.[135] An operational definition given in terms of this two-stage test captures the two mechanisms that bring about depressive symptoms in the above example: the patients who continue to score high on the second test administration have "C-block depression" while the patients whose scores dropped have "C-less depression". Such a concept has the benefits of operational analysis and of mechanistic analysis: being test-based, the characterization is likely to advance intersubjective agreement in diagnosis; being mechanistically grounded, the characterization helps us intervene effectively.

All in all, operationalism has evoked many worries. I have responded to what I consider to be the most pertinent objections to operationalism, arguing that the engineer of operational concepts need not fall prey to these objections. Table 14 summarizes these worries and my responses.

[135] I am ignoring all the relevant ethical consideration just to illustrate what the operationalist's response might be. In the diabetes case, the operationalist might define the subtypes of diabetes in terms of different reactions in a test where insulin and glucose are administered simultaneously to a patient with diabetic symptomatology. A type I patient has a similar reaction as an asymptomatic individual (person who manifests none of the diabetic symptomatology) while type II patient's blood sugar rises radically. This was the test set-up Himsworth (1936) used to show the existence of two different kinds of diabetes.

| Worry | Response |
|---|---|
| Operational definitions are arbitrary. | Operational definitions can be useful and respectful of meanings attached to the target concept. |
| Operationalism leads to proliferation of concepts | Old operational concepts can be replaced by new ones. Sometimes multiple operational definitions are useful. |
| Operationalists are anti-realists (and anti-realism is not tenable). | Operationalism is often motivated by cautiousness, not anti-realism |
| Operationalism fails to explain and enable intervention the way mechanisms do | Operationalism serves the attainment of other epistemic values. It is complementary to mechanistic characterizations of target concepts. |

*Table 14. Worries and responses to worries about operationalism.*

### 6.3 Inferentialism and Respectful Operationalism

#### *6.3.1 Validation Dualism*

Now that we have steered clear from some implausible caricatures of operationalism, I would like to use these insights to support a claim about psychometric validation. My claim is

Validation Dualism: *There are two distinct, defensible approaches to psychometric validation, the respectful operationalist approach and the inferentialist approach*.[136]

Inferentialism and respectful operationalism employ the same psychometric toolbox, which I described in chapter 2: statistical tests of model-fit, correlations between tests and so on.[137] Both engage in conceptual engineering, and both are representing empirical relations. But their use of the psychometric tools, the concepts they engineer and the relations they represent differ radically from each other.

[136] Inferentialism could also be called realism, for reasons that will become clear. This would, however, suggest a contrast with anti-realism, but that is not the correct contrast here. To avoid this association, I shall use the term inferentialism.

[137] I will here outline my proposal in broad brush strokes, and hence cannot go through dual interpretations of the whole psychometric toolbox. It is conceivable that some psychometric methods do not have an interpretation within either respectful operationalism or inferentialism. If that is the case, I would be willing to make my claim more moderate and express it e.g. in terms of "most of psychometric tools".

While inferentialism and respectful operationalism have not been articulated before, they certainly pack insights from existing literature and practice – these have been introduced throughout preceding chapters and sections. To defend Validation Dualism, I will first explicate and defend respectful operationalism and inferentialism separately, and then explicate their relations. Finally, I will explain how failure to distinguish between the two leads to epistemic havoc.

### *6.3.2 Respectful Operationalism*

To understand respectful operationalism, let us first consider a fictional villain: a psychometrician engaged in *disrespectful* operational validation. I will call him *the disrespectful operationalist* for short. The disrespectful operationalist is disrespectful in this sense: he has no interest in what excess meaning depression has beyond his operational definition. For example, he does not care whether people tend to associate depression with anxiety and stress and chemical imbalances in the brain – for all he cares a test of depression might ask about a person's favourite football team. Despite such disrespect for meanings, he might be interested in the techniques that typically go under the banner "validation", for example, tests of reliability and model-fit. But in the disrespectful operationalist programme, these reliability and model-fit checks are not meant to ensure that the measure captures what it is meant to capture – that would not make sense because the disrespectful operationalist, being operationalist, wants to capture what the measure captures. The role of the reliability and model-fit checks is rather to establish that the test, and hence the concept that is defined in terms of the test, has some desirable properties. For example, the disrespectful operationalist might want high reliability coefficients, not because he thinks depression is homogeneous or first factor saturated, but because he wants a test that has this property, i.e. questions that tend to receive consistent responses.

How does the disrespectful operationalist choose which properties the test should have? Being disrespectful, he ignores conceptual intuitions and usage and goes with whatever properties he wants: depression with no association to stress and anxiety, well-being with no relation to physical health, and so on. With the disrespectful operationalist strategy, concepts of e.g. depression can proliferate more or less indefinitely, because nothing

constraints what gets called depression, except the whims of the test constructor. It should go without saying that such an approach is suboptimal.

Let's introduce a different kind of character: a person who also wants an operational definition but is willing to concede that many psychological concepts, depression included, have meaning that goes beyond psychological tests. Let's say she cares about that "extra-operational" meaning, that is, the slippery and unruly associations that manifest in all human practices concerning this thing that is called "depression" (most of us do care about extra-operational meanings, which is why it is cognitively burdensome to make sense of the game disrespectful operationalists play). We might call her *the respectful operationalist*: a test constructor who respects the meanings people attach to depression, but who nonetheless takes her test to *define,* rather than track, a concept of depression.

The test construction process of the respectful operationalist is much more constrained than that of her disrespectful colleague. The extra-operational meaning of depression guides the way the respectful operationalist selects test properties. Perhaps she decides her measure should correlate with measures of stress and anxiety, because people conceive of depression as related to stress and anxiety.[138] Nonetheless, in our everyday meaning apparatus, depression is not the same as anxiety or stress. The depression test should reflect this, the respectful operationalist decides, and this leads her to ask laypeople and depression experts about whether the proposed test questions appear to pertain to depression rather than anxiety and stress.

Extra-operational meanings offer one set of constraints on the respectful operationalist's test construction project. But there is another set of constraints that is largely independent of extra-operational meaning constraints: constraints that pertain to the usability of the test. For example, the test maker may want the test to be quick to administer – hence long questionnaires are excluded. More substantively, the test constructor may want the test to predict events that are associated with the extra-operational meaning of depression, for example, suicide attempts. These considerations of usability form another set of constraints the test developer must take into account.

[138] Relating to anxiety, this is demonstrated e.g. by the fact that HAM-D contains questions about anxiety. See discussions in Kendall and Watson (1989) for more on the relation between anxiety and depression. On stressful life events and depression, see e.g. Kendler et al. (1998).

The respectful operationalist's project is characterized by conceptual engineering through and through: the construction of the test constitutes the construction of the target concept. The extra-operational meaning of the concept, and the aims of the test maker, suggest features the test – and the concept! – should have, and the test maker decides which constraints she wants to take on board and which ones to dismiss. She shuffles back and forth between considering the demands of the extra-operational meaning of depression, empirical study of the properties of the test and revision of test questions, until these align with each other to her satisfaction. To align all these aspects is hard: perhaps laypeople and experts have incompatible conceptions of depression (e.g. some say depression is a chemical imbalance, others say it's not (Lacasse and Leo 2005; France, Lysaker, and Robinson 2007)); perhaps none of the tests she tries fulfil a popular conception of depression (e.g. none of her proposed tests correlates sufficiently with serotonin deficiency); perhaps the test has worse predictive capacities when a particular bit of extra-operational meaning is incorporated in the test (e.g. the ability of the test to predict suicide attempts suffers when "proneness to physical violence" is not included among the items, but physical violence is not core to the extra-operational meaning of depression). Not every part of the extra-operational meaning can be taken on board, not every aim can be perfectly achieved – choices must be made.

When the alignment is to the satisfaction of the test maker, she declares that her test defines a concept of depression. The justification for this declaration is that, when constructing the test, she has been sensitive to the way laypeople and/or experts conceptualize depression. She has taken on board those meanings in so far as they do not undermine prediction and usability. Her test, therefore, defines a concept of depression.

Compare respectful operationalism to what I have previously called construct validation: one starts with a theory of connections between depression and related concepts such as anxiety and stress; checks whether the proposed test of depression correlates with measures of those other concepts in the expected manner; if the proposed test of depression correlates with tests of anxiety and stress, the test is (provisionally, pending further study) accepted; if they do not correlate, either the theory is deemed incorrect and revised, or the test is revised. Unlike construct validation, the respectful operationalist does not start with a theory and proceed to test it. She starts with raw conceptual materials and proceeds by construction and re-construction – if there is theory involved, it is typically a messy folk theory

of depression, or whatever it is she is out to measure.[139] When two tests do not correlate in the expected manner, the respectful operationalist and the construct validator respond in different ways. For example, if a proposed test of depression fails to correlate with a test of anxiety, the respectful operationalist can retain the test, retain the idea that anxiety is related to some useful concepts of depression but propose that the new test defines a concept of depression that is not related to anxiety. Why might she choose that rather than improve the test or revise her theory of depression?

It should be clear by now that in the respectful operationalist approach, validation is about balancing two types of considerations: i) considerations of meanings attached to the target concept, and ii) considerations of the usability of the test. Sometimes meanings contradict each other (e.g. some think of depression in terms of chemical imbalances, others reject this view altogether (France, Lysaker, and Robinson 2007)). At other times, meanings and usability pull the test constructor in opposite directions (e.g. a short test is easier to use but the complexity of meanings associated with depression is better captured by a long test). The navigation of such oppositions requires leaving out some *defensible* connotations of the target concept, which is why conceptual pluralism might be a better option than revision of the test or the background theory. Respectful operationalist validation comes apart from traditional construct validation by replacing theory revision with concept division.

That may sound like a minor difference. The big one is yet to come. The real peculiarity of respectful operationalism is that the numbers psychometric instruments yield are accepted as classificatory, ordinal or quantitative *on the basis of considerations of usability, not on the basis of empirical evidence for the accuracy or correctness of the representation.* This is because the representational claims the respectful operationalist makes are trivially true once the test has been constructed. The operationalist is claiming that her numbers represent the operation she constructed, or more precisely, a concept that is defined solely in terms of the test she constructed. The operationalist test of depression, for

[139] There may also be a "theory of usefulness" involved, that is, some account that justifies the kinds of uses the test maker envisions for her measure. The theory that is lacking is a theory of the test-independent attribute or attributes that drive the test results.

example, is not informative of some test-independent notion of depression. [140] This, recall, is *not* because the operationalist has to believe that there are no real, empirically discoverable patterns that determine the way people score on the test – she does not have to be an anti-realist! – but because she has *not validated* any claims about the representational relation between numbers and whatever underlies score determination. Equality of the intervals between numbers, for example, is just an artefact of the fact that the test yields numbers.

Why then talk about quantitative, ordinal or classificatory numbers, if the usual representational interpretation is replaced by trivially true claims about the test set-up? The only meaning of "interval interpretation" in the operationalist case is that the operations that are associated with interval scales, such as addition and comparisons of differences, are being used. The justification of that usage is that it yields useful results: for example, it could be that the operation of addition is a precondition of successful prediction. Note that it is not appropriate to describe the operationalist by saying that she "assumes interval level properties". This is not appropriate, because assumptions are usually taken to be truth-apt, that is, assumptions can be evaluated in terms of truth and falsity. But that is not the case here: the operationalist cannot be wrong about his scale type, because his claims about scale type are choices, not truth-apt claims. In choosing the scale type, the respectful operationalist merely expresses a preference for using operations such as addition and comparison of differences. The very meaning of scales has been altered in the operationalist project.

It may seem like foul play that respectful operationalists regard claims about scale type true by construction. My sense is that to many laypeople and scientists, measurement connotes representation of something test-independent, some relations or phenomena that exist whether or not anyone measures them.[141] Consequently, scale type

[140] It is worth mentioning an objection here: how can the operationalist incorporate extra-operational meanings into her test without simultaneously gaining better and better knowledge of the test-independent something that brings about test responses? After all, is it not the case that *intensions determine extensions,* that is, meanings determine entities the concept applies to? The answer is that the operationalist's meaning building leaves a lot of room for uncertainty about what drives responses. Correlations between a depression test and an anxiety test, for example, could result from multiple processes, for example, i) test takers want to come across as consistent and therefore answer in a similar manner in both tests, ii) test takers agree that questions across tests have more or less the same meaning but disagree on what that meaning is (Rhemtulla, Borsboom, and Van Bork 2017), or iii) test takers give responses they expect the test administrator wants to hear. Given this, the operationalist can build connections between instruments in accordance with extra-operational meanings but not know much about what underlies score determination.

[141] I am not aware of any empirical work on conceptual intuitions regarding scales and measurement.

assumptions are thought to be truth-apt claims, not something that can be established by stipulation. Because of its peculiar interpretation of scales, respectful operationalism might look like an odd mix of respectfulness and disrespectfulness: while it respects some of the extra-operational meanings attached to a target concept, it disrespects some common meanings attached to scales, and more broadly, meanings attached to the claim that "this test *measures* depression".

Nonetheless, we should not be scornful of respectful operationalism, for two reasons. First, usefulness is useful, that is, if operationalist instruments yield numbers that have, for example, predictive functions, there is nothing wrong with using them for predictive purposes (as long as one steers clear of unfounded non-operational claims of the kind discussed in chapter 4). HAM-D, for example, could be a useful variable in a complex model that predicts remission or suicide attempts, even though HAM-D is very unlikely to represent any defensible, test-independent concept of depression (Bagby et al. 2004; Kramer 2016). Furthermore, the operationalist validation process ensures that the *test itself* has interesting properties, which is why one might be genuinely interested in the ratings *in and of themselves*, regardless of what underlies or drives those responses. Finally, and as mentioned in section 6.2.1 above, operationally validated instruments can also serve intersubjective agreement and other valuable aims test makers have.

Second, even though the numbers a psychometric instrument yields do not have a test-independent representational interpretation in virtue of the operationalist validation process, it *might* turn out that the operationalist validation provides a stepping stone to such a representation. For example, perhaps it turns out that, although it was not the intention of the operationalist, the predictive capacities she has built into her test (and her concept) are due to the instrument (approximately) representing an interesting, non-operationally defined concept of depression. If that is the case, the operational project might turn into an inferential one.

### *6.3.3 Inferentialism*

Consider a psychometric instrument that tests whether a person is or is not in remission from depression. The typical approach to such a classification is to define a threshold score such that individuals who score lower than the threshold on the proposed test

are classified as in remission, while those who score higher than the threshold are not in remission. For example, for the HAM-D the threshold defining remission is 7 (Frank et al. 1991).[142] How would such a classificatory usage be validated, if the intention is to capture a non-operationally defined target concept? In other words, what does validation look like, if the test needs to allow inferences that go beyond the test, to a feature of the test takers that drives or determines the way they respond on the test?

Just like the operationalist psychometrician, the inferentialist psychometrician must consider meanings attached to the target concept and the usability of the test. If the goal is to validate an instrument that classifies patients according to whether or not they are in remission from depression, the concepts of remission and non-remission will have to resonate with at least some of the meanings experts and/or laypeople attach to these concepts. Likewise, the usability of the instrument has to be such that the instrument fits the intended context: not too many questions if speed is of the essence, and so on for other usability considerations. So far, very similar to the game the respectful operationalist plays.

The crucial difference is that the inferentialist must consider a third set of constraints. The third constraint is that the instrument must yield numerical representations that *point beyond the instrument itself*. In the depression case, the inferentialist has to say something more substantial than that remission is defined by HAM-D scores below 7 and that non-remission is defined by scores above 7. Rather, the concept of remission should denote *that which brings about* a HAM-D score below 7, while non-remission should denote *that which brings about* a HAM-D score above 7. For example, non-remission might denote a dysfunctional neurotransmitter mechanism, which brings about the symptoms that get reported on the HAM-D score, while remission denotes a normally functioning neurotransmitter system, which brings about the asymptomatic state that gets reported via the HAM-D. More likely, remission might denote a wholesomely described state: a sense of well-being and clarity, robustness in the face of set-backs, hopefulness, physical well-being, ability to take care of oneself and others, and so on and so forth. Non-remission could be characterized in contrary terms as the state that brings about a high score.

[142] In Frank et al. (1991), the period the below 7 score must be sustained to count as full remission is 2 weeks to 6 months.

How does one validate such a non-operational classification? Like the respectful operationalist, the inferentialist uses the psychometric tools I described in chapter 2. But in the inferentialist project, these tools are used, not only to engineer concepts, but also *to confirm theories and hypotheses about what determines scores.* For example, the inferentialist could use correlations between various administrations of HAM-D to infer, not just the stability of scoring, but more importantly the stability of the hypothesized state that drives that scoring. This will require more than just reporting test-retest correlations: there will have to be a holistic argument to infer stability of the target state from stability of the test results. To give another example: a correlation between HAM-D and a measure of anxiety ideally allows the inferentialist to infer, not (just) that his target concept coheres with extra-operational meanings attached to depression, but that the state that determines the scoring is the state he hypothesized it to be (say a dysfunctional neurotransmitter system). This again requires argumentation, because correlations between tests do not automatically mean that the test tracks the test-independent state or process of interest.

The validation the inferentialist goes through is laborious and epistemically risky. He can easily be wrong about the representational capacities of his numbers. For example, perhaps his measure fails to classify people in terms of the target concept. Perhaps what determines people's scores on a test is not what the inferentialist thought determines them. Perhaps what determines scores varies radically from case to case. The reward for facing these challenges is epistemically strong claims about a test-independent attribute.

### *6.3.4 A dual interpretation of the psychometric toolbox*

| **Respectful Operationalism** | **Inferentialism** |
|---|---|
| Represents test responses | Represents determinants of test responses |
| Representation by construction | Representation by hypothesis testing |
| Engineering operational concepts | Engineering non-operational concepts |
| Engineering constraint by extra-operational meaning and usability | Engineering constraint by the nature of test-independent reality, extra-operational meaning and usability |

*Table 15. Comparison of respectful operationalism and inferentialism.*

Table 15 presents a comparison of key features of respectful operationalism and inferentialism. Both kinds of psychometricians, the operationalist and the inferentialist, are engaged in some kind of representational activity: the operationalist psychometrician

represents test responses while the non-operationalist psychometrician represents that which drives test responses. The difference is that the operationalist gets her representation by construction (i.e. whenever one has a test, one has a representation) while the non-operationalist's representation is at the mercy of test-independent reality, which he approaches via rigorous hypothesis testing.

Both the operationalist and the inferentialist are doing conceptual engineering: they are forming the test and the target concept simultaneously in light of considerations of extra-operational meanings and test usability. The difference is that only the inferentialist is constrained by push-back from test-independent reality (as I called it in chapter 5).

Both the respectful operationalist and the inferentialist are using the same tools described in chapter 2. But they use them in different ways and give radically different interpretations to the results. To go through all the tools thoroughly would take a chapter (or a dissertation) of its own. But let us discuss a few of the most common tools, to get a flavour of the contrasting interpretations these tools can have.

Start with an example we already have some familiarity with: checks of correlations between tests. In the inferentialist approach, the interpretation of inter-test correlations proceeds along the lines of classical statements of construct validation (section 2.6.1). In other words, correlations are assessed in terms of a theory and used to make inferences to a test-independent state or process: if correlations cohere with theoretical expectations, there is reason to believe that the measure tracks the state or process of interest. In the operationalist approach, the use of inter-test correlations is two-fold. Firstly, the operationalist can utilize the new test to predict scores on other, already established tests, if she knows the relationship between the two tests. This way of thinking about inter-test correlations connotes classical notions of criterion and predictive validity (section 2.6.2). Secondly, the operationalist can use inter-test correlations to justify the term she applies to her operational concept.

Move on to another example, model-fitting exercises in the Item Response Theory paradigm. Such model-fitting has a straightforward inferentialist interpretation. IRT models specify the functional relationship between observed responses and attributes, where attributes such as ability are hypothesized to exist independent of the testing. Tests of model fit can therefore be used to confirm or falsify hypotheses about the determinants of the

observed scores, in other words, to support inferences to the non-observed attributes of interest. The operationalist interpretation is a little less evident.[143] One option is to say that "ability levels" – that is, the things depicted on the x-axes in Figures 5 and 6 – are compact expressions of test response patterns, i.e. a given "ability level" summarizes probabilities of correct response to each item. In other words, "ability level" is a summary of test response patterns, not a real property of an individual that brings about that test behaviour. Another option for the operationalist is to think of "ability levels" as *place-holders* for as-yet-untheorized and "uninferred", real aspects of the test and the test takers. In both of these options, the typical IRT terminology is misleading and should probably be accompanied by declarations of operationalism.

The above is far from an all-inclusive account of how the inferentialists and the operationalists can make use of the psychometric toolbox. It could be that there are some psychometric tools that do not have a respectful operationalist interpretation, and that some tools do not have an inferentialist interpretation. These details are important, but not our concern in this chapter. My point is just that both approaches can make use of common psychometric methods.

### 6.4 Mixing the mismatching

I believe that inferentialism and respectful operationalism are both defensible ways of pursuing psychometrics. They do not, however, mix well. In particular, one should not validate operationally and make inferential claims – that is, one should not walk the operationalist walk and talk the inferentialist talk. The reason is simply that operationalist validation does not provide epistemic access to the kinds of claims inferentialists make. To know how people respond to questions about depression, however nicely formulated those questions are, is far from knowing about people's feelings, thoughts, capacities, brains, genes and other test-independent features that drive the way people respond. To pretend otherwise

[143] Borsboom, Mellenbergh, and van Heerden (2003) write: "We conclude that operationalism and latent variable theory are fundamentally incompatible." I think they mean that common interpretations of latent variable theory, as manifested in the usage of e.g. IRT models, are incompatible with some forms of operationalism. But as we have seen, there are many kinds of operationalist approaches, just like there is lots of room for variety in applications of IRT models. I therefore think that Borsboom et al.'s conclusion does not undermine the point I am making here.

– as someone using operationalist validation for inferentialist claims would – is to be confused or lying.

If we allowed inferentialist claims to be made on the basis of an operationalist validation, a whole lot of epistemic havoc and outright non-sense would follow. It is one thing to be justified in claiming that in an anti-depressant trial, average scores on a test lowered more in the treatment group than the control group. It is another thing to be able to say that that score change represents a change in chemical balances in the patients' brain, rather than, say a change in patients' average proneness to lying, or a mixed bag of other non-chemical changes. Claims about the effectiveness of a drug are more convincing and useful, when one knows what test-independent thing drives the difference between control and treatment groups. Currently there is not much agreement on what score changes on common depression measures mean in concrete, test-independent terms (e.g. Leucht et al. 2013). It is therefore not surprising that the effectiveness and usefulness of anti-depressants is contested (see section 5.3.2).

As another example of how mixing operationalism and inferentialism leads to confusion, consider standards for establishing scale types. When is a researcher warranted to say that she has a valid, quantitative representation? To the inferentialist, claims about scale types are genuine epistemic achievements that establish patterns in reality. To the operationalist, claims about scales are stipulations or expressions of preference, which state what operations are applied to numerical data. Making inferentialist claims on the basis of an operationally validated instrument quickly leads to inconsistency. Proponents of RTM sometimes accuse S.S. Stevens of such inconsistency. For example, R. Duncan Luce (1997b) once wrote that "it is doubtful" as to whether Stevens understood the meaning of interval or ratio scale representation. This is remarkable, because as Luce knew, Stevens was the first to recognize and emphasize the existence of different scale types. According to Luce and his colleagues, Stevens treatment of scales was ambiguous. The ambiguity is arguably because Stevens' usage of scales had both operationalist and inferentialist flavours (see Krantz et al. 1971, section 1.2.3 and fn. 2).[144]

[144] This is my reading of the worries Luce and others had. Luce and his colleagues did not use the term "inferentialism" and did not provide an extensive analysis of the way operationalism showed up in Stevens' treatment of scale types.

To give one more example of mix ups, consider debates about measurability. Is well-being – a complex, personal and hard-to-communicate phenomenon – measurable? To many people, this is a genuine question: only a thorough investigation will tell whether well-being is measurable or non-measurable (e.g. Hausman 2015 thinks well-being is not measurable). However, in the operationalist project, everything is measurable. A test can always be constructed, and the target concept, say well-being, defined in terms of that test. Sure, the respectful operationalist must incorporate some extra-operational meanings by checking that her measure of well-being relates to measures of illness, happiness and other such well-being relevant concepts in a plausible way. But still, the operationalist's constraints are so weak that she can be respectful and still measure pretty much anything. Debates about measurability become opaque, if operationalists fail to make their peculiar position known.

| | **Respectful operationalism (RO)** | **Inferentialism** |
|---|---|---|
| **Primary epistemic benefit** | Intersubjective agreement, "safety" | Understanding, explanation, intervention |
| **Scale type** | Established by stipulation | Established by evidence and inference |
| **Measurability** | Non-issue: everything is measurable in the RO sense | Measurability is a genuine question to be grappled with |

*Table 16. Differences between inferentialism and respectful operationalism.*

Table 16 summarizes differences between respectful operationalism and inferentialism. I argued above that these are differences we must be aware of in order to not mislead each other, in order to not talk past each other. The worry about epistemic havoc may appear purely hypothetical – but it is not. I think that the diagnosis of psychometrics I presented in chapter 4 can be recast in terms of an unfruitful mixture of operationalism and inferentialism. Recall that I argued that, on the one hand assumptions about quantitative representation of non-operationally defined concepts are common, and that on the other hand, the typical validation process does not warrant claims about quantitative, non-operational representation. A clear instance of mixing incompatible views.

It also seems to me that much of the messiness of the psychometric literature (which we reviewed in chapter 2) can be explained in terms of a failure to distinguish inferentialist ideas from operationalist ideas. Consider, firstly, the debate about true scores (see 2.4.1 and 2.5.4): is true score the test taker's true standing on some test-independent

target attribute or the expected score on infinite administrations of the same test. The former response clearly manifests the inferentialist ideal, while the latter approach comes closer to respectful operationalism. In this debate, inferentialists and operationalists are talking about very different things with the same terminology, that is, the terminology of true scores and Classical test theory.

As a second example, consider the confusingly varying ideas about reliability in the psychometric literature: some see reliability coefficients as pertaining to accuracy, while others conceptualize them as indicators of repeatability and representativeness (see 2.5, especially 2.5.4). This divergence can be made sense of in terms of an implicit division to inferentialists and respectful operationalists – "implicit", because divergent views of what reliability stands for are not indexed with declarations of operationalist or inferentialist commitments. An operationalist could easily use test-retest reliability to claim that test-responses are repeatable in certain kinds of test set-ups, while an inferentialist would want to go further and make claims about the accuracy of the test in tracking the test-independent state or process of interest.[145] It is no wonder that ideas about reliability become hopelessly entangled, if psychometricians fail to distinguish respectful operationalism from inferentialism.

I have argued that mixing operationalist and inferentialist ideas is a bad idea. Nonetheless, mix ups seem to occur – why? One plausible reason, in my view, is the partly undeserved disrepute of operationalism. Operationalism being demonized, people whose activities cohere with the respectful operationalist paradigm cannot voice that commitment without risking rejection of their work. Sociologically speaking, it may be the most viable strategy to talk the inferentialist talk but walk the operationalist walk. Furthermore, the mix of inferential talk and operational walk is rhetorically incredibly powerful: the operational validation process is epistemically undemanding, and the inferential claims are strong. If a researcher manages to remain ignorant of (or ignore) the fact that her validation does not warrant the claims she makes, it is hardly surprising she will enjoy the combination of ease and strength the inferentialist-operationalist mix seemingly affords. But as the multiple

[145] Similarly, an operationalist could make sense of parallel test reliability in terms of representativeness: how representative is a test taker's observed score of their success on repeated administrations of the test?

examples here discussed show, we should resist the temptation to walk the operationalist walk and talk the inferentialist talk.

### 6.5 Validation Dualism

In this chapter I have first provided a historical overview of operationalist ideas, and then defended the usefulness of engineering operational concepts. Having thus undermined some common prejudices against operationalism, I argued for a normative thesis called Validation Dualism, which states that

*There are two distinct, defensible approaches to psychometric validation, the respectful operationalist approach and the inferentialist approach*.

The two approaches validate psychometric tests for different purposes and therefore make different use of the psychometric toolbox. I argued that both inferentialism and respectful operationalism are valuable approaches, but that confusing the two leads to epistemic havoc.

# 7. Conclusion

The constructor of a psychological test has to navigate two apparently opposing aims: depth of valid inferences and intersubjective agreement on those inferences. It is easiest to reach agreement on that which is simple, superficial and standardized. But what we are usually most interested in is that which is complex, deep and hard-to-control.

Consider a psychological test that has captured the popular imagination: *the Rorschach inkblot test*. The test, where subjects are shown ten inkblots and asked to say what they see in them, is largely in disrepute among contemporary psychologists. But throughout the 20th century, the Rorschach was applied in contexts as varied as diagnosis of mental illness, personality assessment and evaluation of criminal responsibility (Searls 2017). It takes a whole lot of expertise to administer the test and interpret the results. To arrive at a mental diagnosis, for example, the psychologist has to consider and juxtapose aspects such as what the subject sees (a bat? a body?), whether the subject focuses on details or the image as a whole, how many things the subject sees, how much the subject focuses on colour versus form of the image, and so on and so forth. There are guidelines for how to incorporate the various considerations, but in its original usage, the diagnosis is the result of the expertise of the examiner, not the application of an unambiguous rulebook.

As the Rorschach test gained popularity around the mid-20th century, it became apparent that the complexity of the test prevented its large-scale application. Thus, when the U.S. military needed large-scale psychological testing during the Second World War, psychologist Molly Harrower responded by streamlining the original Rorschach into a multiple-choice test. In this version, the subject picked one of pre-described answers for each inkblot, and the answers were scored according to a top-secret answer key specifying “good” and “bad” answers to each inkblot. The multiple-choice format, together with an unambiguous scoring key, allowed for efficient administration, ensured agreement across interpreters, and resulted in simple “yes or no”-diagnoses about whether a person was fit for military service. As Harrower herself freely admitted, these efficiency gains came at the cost of the depth of interpretation and explanation that the original Rorschach was meant to provide. But the standardized, rule-governed versions were what lived on, with multiple systems provided and psychometrically tested to vouch for their validity (Searls 2017).

Today psychological science is dominated by standardized tests, more akin to Harrower's rule-governed approach than Rorschach's original, open-ended approach. These standardized tests gain their authority from hard numbers and statistics, not from subjective interpretations by experts. Sure, to construct and validate a psychometric test requires technical prowess and expertise in statistical techniques. But such expertise has an air of objectivity that the inkblot experts lack. Even though most people do not know how psychometricians validate tests, the validity of contemporary psychometric tests comes across as something that anyone could *in principle* check, given enough data and training in statistics.

My first argument in this dissertation suggests that a non-expert who actually carried out such a check would be disappointed, particularly with regard to psychologists' quantitative claims. I have argued that there is a mismatch between i) common ideas of what psychometric tests are valid for, and ii) the kind of claims common psychometric validation techniques are able to warrant. Psychometric instruments are typically thought to yield a quantitative representation of a target concept, where that target concept is thought to stand for a feature or aspect that exists independent of a test. For example, measures of depression are expected to give quantitative information about depression, where the concept of depression is not synonymous with test responses – rather, depression is that which underlies test responses. But I argued that results from tests of reliability, validity and model-fit, as they are typically reported, do not warrant claims about quantitative representation of such non-operationally characterized attributes. Simply reporting favourable numerical results does not suffice for such inferences. What is needed is a more holistic approach, where numerical results of model-fit, reliability and validity are interpreted *relative to each other and from the perspective of quantitative representation*.

Although my conclusion in the first part of the dissertation was sceptical of the *claims* made on the basis of psychometric measurement, it does not undermine the usefulness of the *tools* psychometricians utilize. In the second part of this thesis I argued that to avoid the above-described problem, psychometricians technical prowess could be complemented with a two-pronged interpretation of the *implications* of psychometric checks of reliability, validity and model-fit. I proposed that the psychometric toolbox can be legitimately used for two kinds for projects, each of which pertains to a particular type of concept. Firstly, the respectful operationalist project, which establishes representations of respectfully

engineered operational concepts. Secondly, the inferentialist project, which establishes representations of non-operationally characterized target concepts. I argued that this distinction should be applied sharply and consistently. Specifically, whenever the psychometric toolbox is applied to establish representations of operational concepts, one should never make unreflective inferences to non-operational concepts.

The operationalist and the non-operationalist approaches have distinct merits, matching roughly to the two aims outlined in the beginning of this chapter. The respectful operationalist project is epistemically less challenging, because it operates on the superficial level of test responses. The benefit of remaining on the level of the operation is ease of reaching intersubjective agreement, even in large-scale studies. The inferentialist project is epistemically much riskier, because inferences to the determinants of test-responses can easily go wrong. With lots of room for interpretation, the likelihood of intersubjective agreement is diminished. The payoff of taking these risks is capacity to explain and intervene more effectively.

# Bibliography


Alexandrova, A. 2016. “Can the Science of Well-Being Be Objective?” *The British Journal for the Philosophy of Science* 69 (2): 421–445.

———. 2017. *A Philosophy for the Science of Well-Being*. Oxford: Oxford University Press.

Alexandrova, A., and D. M. Haybron. 2016. “Is Construct Validation Valid?” *Philosophy of Science* 83 (5): 1098–1109.

Andrich, D. 1988. *Rasch Models for Measurement*. Newbury Park, CA: Sage.

———. 2004. “Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms?” *Medical Care* 42 (1, Supplement: Applications of Rasch Analysis in Health Care): 17–116.

Angner, E. 2009. “Subjective Measures of Well-Being: Philosophical Perspectives.” In *The Oxford Handbook of Philosophy of Economics*, edited by Don Ross and Harold Kincaid, 560–579. Oxford: Oxford University Press.

———. 2011. “Current Trends in Welfare Measurement.” In *The Elgar Companion to Recent Economic Methodology*, edited by John B. Davis and D. Wade Hands. Northampton: Edward Elgar.

———. 2013. “Is It Possible to Measure Happiness?” *European Journal for Philosophy of Science* 3 (2): 221–40.

Anthoine, E., L. Moret, A. Regnault, V. Ronique, S. Bille, and J-B Hardouin. 2011. “Sample Size Used to Validate a Scale: A Review of Publications on Newly-Developed Patient Reported Outcomes Measures.” *Health and Quality of Life Outcomes,* 12 (2).

Bagby, R. M., A. G. Ryder, D.R. Schuller, and M. B. Marshall. 2004. “The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight?” *American Journal of Psychiatry* 161 (12): 2163–77.

Beck, A. T, and B. A. Alford. 2009. *Depression: Causes and Treatment.* Philadelphia: University of Pennsylvania Press.

Benjamin, D. J., O. Heffetz, M.S. Kimball, and N. Szembrot. 2014. “Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference.” *American Economic Review* 104 (9): 2698–2735.

Bergen, B. K. 2012. *Louder than Words: The New Science of How the Mind Makes Meaning*. New York, NY: Basic Books.

Bickhard, M. H. 2001. “The Tragedy of Operationalism.” *Theory & Psychology* 11 (1): 35–44.

Bird, A., and E. Tobin. 2018. “Natural Kinds.” *The Stanford Encyclopedia of Philosophy, (Spring 2018 Edition)*.

Blanton, H., and J. Jaccard. 2006. “Arbitrary Metrics in Psychology.” *American Psychologist* 61 (1): 27–41.

Bond, T., and C. M. Fox. 2001. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah NJ: Lawrence Erlbaum.

Booth, M. 2015. “The Danish Don’t Have the Secret to Happiness.” *Atlantic*, 2015.

Borgatta, E. F., and G.W. Bohrnstedt. 1980. “Level of Measurement.” *Sociological Methods & Research* 9 (2): 147–60.

Borsboom, D. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D., and G. Mellenbergh. 2004. “Why Psychometrics Is Not Pathological Science.” *Theory & Psychology* 14 (1): 105–20.

Borsboom, D., G. J. Mellenbergh, and J. van Heerden. 2003. “The Theoretical Status of Latent Variables.” *Psychological Review* 110 (2): 203–19.

Borsboom, D., and A. Scholten. 2008. “The Rasch Model and Conjoint Measurement Theory from the Perspective of Psychometrics.” *Theory & Psychology* 18 (1): 111–17.

Bridgman, P. W. 1927. *The Logic of Modern Physics*. New York: Macmillan.

———. 1959. *The Way Things Are*. Cambridge, Mass.: Harvard University Press.

Brigandt, I. 2011. “Natural Kinds and Concepts: A Pragmatist and Methodologically Naturalistic Account.” In *Pragmatism, Science and Naturalism*, edited by J. Knowles and H. Rydenfelt, 171–96. Frankfurt am Main: Peter Lang Publishing.

Brown, J. F. 1934. “A Methodological Consideration of the Problem of Psychometrics.” *Erkenntnis* 46 (1): 46–61.

Brun, G. 2016. “Explication as a Method of Conceptual Re-Engineering.” *Erkenntnis* 81 (6): 1211–41.

Burgess, A., and D. Plunkett. 2013. “Conceptual Ethics I.” *Philosophy Compass* 8 (12): 1091–1101.

Campbell, D. T., and D. W. Fiske. 1959. “Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.” *Psychological Bulletin* 56 (2): 81–105.

Campbell, S. M. 2016. “The Concept of Well-Being.” In *Routledge Handbook of Philosophy of Well-Being*, edited by G. Fletcher. Abingdon: Routledge.

Cappelen, H. 2012. *Philosophy Without Intuitions*. Oxford: Oxford University Press.

———. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Carmody, T. J, A. J. Rush, I. Bernstein, D. Warden, S. Brannan, D. Burnham, A. Woo, and M. H. Trivedi. 2006. “The Montgomery Asberg and the Hamilton Ratings of Depression: A Comparison of Measures.” *European Neuropsychopharmacology: The Journal of the European College of Neuropsychopharmacology* 16 (8): 601–11.

Carnap, R. 1934. “On the Character of Philosophic Problems.” *Philosophy of Science* 1 (1): 5–19.

———. 1950a. “Empiricism, Semantics, and Ontology.” *Revue Internationale de Philosophie* 4 (11): 20–40.

———. 1950b. *Logical Foundations of Probability.* Chicago: University of Chicago Press.

———. 1955. “Meaning and Synonymy in Natural Languages.” *Philosophical Studies* 6 (3): 33–47.

Casey, B. M., D. D. McIntire, and K. J. Leveno. 2001. “The Continuing Value of the Apgar Score for the Assessment of Newborn Infants.” *New England Journal of Medicine* 344 (7): 467–71.

Chalmers, D. 2014. “Intuitions in Philosophy: A Minimal Defense.” *Philsophical Studies* 171 (3): 535–544.

Champlain, A. De. 2010. “A Primer on Classical Test Theory and Item Response Theory for Assessments in Medical Education.” *Medical Education* 44 (1): 109–17.

Chang, H. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

———. 2009. “Operationalism.” *The Stanford Encyclopedia of Philosophy (Fall 2009 Edition)*.

———. 2017a. “Epistemic Iteration and Natural Kinds: Realism and Pluralism in Taxonomy.” In *Issues in Psychiatry IV: Classification of Psychiatric Illnesses*, edited by Kenneth Kendler and Josef Parnas, 229–45. Oxford: Oxford University Press.

———. 2017b. “Operationalism: Old Lessons and New Challenges.” In *Reasoning in Measurement*, edited by Nicola Mößner and Alfred Nordmann, 25–38. London and New

York: Routledge.

Christensen, K., A. M. Herskind, and J. W. Vaupel. 2006. “Why Danes Are Smug: Comparative Study of Life Satisfaction in the European Union.” *BMJ (Clinical Research Ed.)* 333 (7582): 1289–91.

Cliff, N. 1989. “Ordinal Consistency and Ordinal True Scores” 54 (1): 75–91.

———. 1992. “Abstract Measurement Theory and the Revolution That Never Happened.” *Psychological Science* 3 (3): 186–90.

Coombs, C. H. 1950. “Psychological Scaling without a Unit of Measurement.” *Psychological Review* 57 (3): 145–58.

Cooper, R. 2002. “Disease.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 33 (2): 263–82.

Cortina, J. M. 1993. “What Is Coefficient Alpha? An Examination of Theory and Applications.” *Journal of Applied Psychology* 78 (1): 98–104.

Cronbach, L. J. 1947. “Test ‘Reliability’: Its Meaning and Determination.” *Psychometrika* 12 (1): 1–16.

———. 1951. “Coefficient Alpha and the Internal Structure of Tests*.” *Psychometrika* 16 (3): 297–334.

Cronbach, L. J., and P. E. Meehl. 1955. “Construct Validity in Psychological Tests.” *Psychological Bulletin* 52 (4): 281–302.

Cronbach, L. J., and R. J. Shavelson. 2004. “My Current Thoughts on Coefficient Alpha and Successor Procedures.” *Educational and Psychological Measurement* 64 (3): 391–418.

Cronbach, L.J., N. Rajaratnam, and G. C. Gleser. 1963. “Theory of Generalizability: A Liberalization of Reliability Theory†.” *British Journal of Statistical Psychology* 16 (2): 137–63.

Decoene, S., P. Onghena, and R. Janssen. 1995. “Representationalism under Attack.” *Journal of Mathematical Psychology* 39 (2): 234–42.

Diener, E. 1984. “Subjective Well-Being.” *Psychological Bulletin* 95 (3): 542–75.

Diener, E., R. A. Emmons, R. J. Larsen, and S. Griffin. 1985. “The Satisfaction with Life Scale.” *Journal of Personality Assessment* 49 (1): 71–75.

Diener, E., J. J. Sapyta, and E. Suh. 1998. “Subjective Well-Being Is Essential to Well-Being.” *Psychological Inquiry* 9 (1): 33–37.

Dutilh Novaes, C. 2018. “Carnapian Explication and Ameliorative Analysis: A Systematic Comparison.” *Synthese*, February, 1–24.

Dutilh Novaes, C., and E. Reck. 2017. “Carnapian Explication, Formalisms as Cognitive Tools, and the Paradox of Adequate Formalization.” *Synthese* 194 (1): 195–215.

Eklund, M. 2015. “Intuitions, Conceptual Engineering, and Conceptual Fixed Points.” In *The Palgrave Handbook of Philosophical Methods*, edited by C. Daly, 363–85. London: Palgrave Macmillan.

Embretson, S., and S. Reise. 2000. *Item Response Theory for Psychologists*. Mahwah NJ: Lawrence Erlbaum Associates Publishers.

Emre, M. 2018. *What’s Your Type?: The Strange History of Myers-Briggs and the Birth of Personality Testing*. HarperCollins.

Ereshefsky, M. 2018. “Natural Kinds, Mind Independence, and Defeasibility.” *Philosophy of Science* n/a (n/a).

Feest, U. 2005. “Operationism in Psychology: What the Debate Is about, What the Debate Should Be About.” *Journal of the History of the Behavioral Sciences* 41 (2): 131–49.

Feigl, H. 1950. “Existential Hypotheses. Realistic versus Phenomenalistic Interpretations.” *Philosophy of Science* 17 (1): 35–62.

Fischer, G. H., and I. W. Molenaar. 1995. *Rasch Models*. New York, NY: Springer New York.

Fiske, D. W. 1971. *Measuring the Concepts of Personality*. Chicago: Aldine Pub. Co.

Fraassen, B. van. 2008. *Scientific Representation*. Oxford: Oxford University Press.

France, C. M., P.H. Lysaker, and R. P. Robinson. 2007. “The Chemical Imbalance Explanation for Depression: Origins, Lay Endorsement, and Clinical Implications.” *Professional Psychology: Research and Practice* 38 (4): 411–20.

Frank, E., R. F. Prien, R. B. Jarrett, M. B. Keller, D. J. Kupfer, P. W. Lavori, A. J. Rush, and M. M. Weissman. 1991. “Conceptualization and Rationale for Consensus Definitions of Terms in Major Depressive Disorder. Remission, Recovery, Relapse, and Recurrence.” *Archives of General Psychiatry* 48 (9): 851–55.

Frege, G. 1892. “On Sense and Reference.” In *Translations from the Philosophical Writings of Gottlob Frege (1980)*, edited by P. Geach and M. Black. Oxford: Blackwell.

French, C. F. 2015. “Philosophy as Conceptual Engineering: Inductive Logic in Rudolf Carnap’s Scientific Philosophy.” University of British Columbia.

Frigg, R., and J. Nguyen. 2016. “Scientific Representation.” *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*.

Fujiwara, D., and R. Campbell. 2011. “Valuation Techniques for Social Cost-Benefit Analysis: Stated Preference, Revealed Preference and Subjective Well-Being Approaches: A Discussion of the Current Issues.” HM Treasury and Department for Work and Pension, UK.

Furr, R. M. 2011. *Scale Construction and Psychometrics for Social and Personality Psychology*. London: Sage.

Galluzzo, G, and M. J. Loux. 2015. *Problem of Universals in Contemporary Philosophy*. Cambridge: Cambridge University Press.

García-Batista, Z. E., K. Guerra-Peña, A. Cano-Vindel, S. X. Herrera-Martínez, and L. A. Medrano. 2018. “Validity and Reliability of the Beck Depression Inventory (BDI-II) in General and Hospital Population of Dominican Republic.” *PLOS ONE* 13 (6): e0199750.

Gerring, G. 1999. “What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences.” *Polity* 31 (3): 357–93.

Goertz, G. 2006. *Social Science Concepts: A User’s Guide*. Princeton, NJ: Princeton University Press.

Goldman, A. I. 1976. “Discrimination and Perceptual Knowledge.” *Journal of Philosophy* 73: 771–91.

Goodman, N. 1954. *Fact, Fiction and Forecast*. London: University of London.

Green, C. D. 1992. “Of Immortal Mythological Beasts.” *Theory & Psychology* 2 (3): 291–320.

Grice, H. P. 1989. *Studies in the Way of Words*. Cambridge MA: Harvard University Press.

Gulliksen, H. 1950. *Theory of Mental Tests*. New York: Wiley.

Hacking, I. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

———. 1999. *The Social Construction of What?* Cambridge, Mass.: Harvard University Press.

Hall, B. K. 1999. “The Paradoxical Platypus.” *BioScience* 49 (3): 211–18.

Hambleton, R. K., H. Swaminathan, and H. J. Rogers. 1991. *Fundamentals of Item Response Theory*. Thousand Oaks: Sage Publications.

Haslanger, S. A. 2000. “Gender and Race: (What) Are They? (What) Do We Want Them To Be?” *Noûs* 34 (1): 31–55.

———. 2008. “A Social Constructionist Analysis of Race.” In *Revisiting Race in the Genomic Age*, edited by B. Koenig, S. Lee, and S. Richardson. Piscataway, N.J.: Rutgers University Press.

———. 2012. *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press.

Hausman, D. M. 2015. *Valuing Health*. Oxford: Oxford University Press.

Haybron, D. M. 2003. “What Do We Want from a Theory of Happiness?” *Metaphilosophy* 34 (3): 305–29.

Haybron, D. M., and A. Alexandrova. 2013. “Paternalism in Economics.” In *Paternalism. Theory and Practice.*, edited by C. Coons and M. Weber. Cambridge: Cambridge University Press.

Haybron, D. M., and V. Tiberius. 2015. “Well-Being Policy: What Standard of Well-Being?” *Journal of the American Philosophical Association* 1 (04): 712–33.

Heilmann, C. 2015. “A New Interpretation of the Representational Theory of Measurement.” *Philosophy of Science* 82 (5): 787–97.

Helliwell, J., R. Layard, and J. Sachs. 2017. “World Happiness Report 2017.” New York.

Hempel, C. G. 1954. “A Logical Appraisal of Operationism.” *The Scientific Monthly* 79: 215–20.

———. 1966. *Philosophy of Natural Science*. Englewood Cliffs, N.J: Prentice-Hall.

Hieronymus, F., J. F. Emilsson, S. Nilsson, and E. Eriksson. 2016. “Consistent Superiority of Selective Serotonin Reuptake Inhibitors over Placebo in Reducing Depressed Mood in Patients with Major Depression.” *Molecular Psychiatry* 21 (4): 523–30.

Himsworth, H.P. 1936. “Diabetes Mellitus: Its Differentiation into Insulin-Sensitive and Insulin-Insensitive Types.” *The Lancet* 227 (5864): 127–30.

Hintikka, J. 1999. “The Emperor’s New Intuitions.” *Journal of Philosophy* 96 (3): 127–47.

Hood. 2008. “Latent Variable Realism in Psychometrics.” Bloomington: Indiana University.

Horton, M., I. Marais, and K. B. Christensen. 2013. “Dimensionality.” In *Rasch Models in Health*, edited by K. B. Christensen, S. Kreiner, and M. Mesbah, 137–58. Hoboken, NJ: John Wiley & Sons, Inc.

Howell, D. C. 2010. *Statistical Methods for Psychology*. Belmont, CA: Thomson Wadsworth.

Hull, D. L. 1968. “The Operational Imperative: Sense and Nonsense in Operationism.” *Systematic Zoology* 17 (4): 438–57.

Illari, P. M. 2011. “Mechanistic Evidence: Disambiguating the Russo–Williamson Thesis.”

*International Studies in the Philosophy of Science* 25 (2): 139–57.

Isaac, A. M. 2013. "Objective Similarity and Mental Representation." *Australasian Journal of Philosophy* 91 (4): 683–704.

Jackson, F. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

John, O. P., and V. Benet-Martinez. 2000. "Measurement: Reliability, Construct Validation, and Scale Construction." In *Handbook of Research Methods in Social and Personality Psychology*, edited by H. T. Reis and C. M. Judd, 339–69. NY, US: Cambridge University Press.

Jones, L.V., and D. Thissen. 2006. "A History and Overview of Psychometrics." In *Handbook of Statistics*, edited by S.R. Rao and S. Sinharay, Volume 26, 1–27.

Judd, C., and G. McClelland. 1998. "Measurement." In *Handbook of Social Psychology*, edited by S. Fiske, D. Gilbert, and G. Lindzey, 4th editio. Boston: McGraw-Hill.

Keats, J. A. 1967. "Test Theory." *Annual Review of Psychology* 18 (1): 217–38.

Keller, M. B., N. D. Ryan, M. Strober, R. G. Klein, S. P. Kutcher, B. Birmaher, O. R. Hagino, et al. 2001. "Efficacy of Paroxetine in the Treatment of Adolescent Major Depression: A Randomized, Controlled Trial." *Journal of the American Academy of Child and Adolescent Psychiatry* 40 (7): 762–72.

Kendall, P. C., and D. Watson. 1989. *Anxiety and Depression: Distinctive and Overlapping Features*. San Diego: Academic Press.

Kendler, K, S., L. M. Karkowski, and C. A. Prescott. 1998. "Stressful Life Events and Major Depression: Risk Period, Long-Term Contextual Threat, and Diagnostic Specificity." *The Journal of Nervous and Mental Disease* 186 (11): 661–69.

Khan, A., H. A. Warner, and W.A. Brown. 2000. "Symptom Reduction and Suicide Risk in Patients Treated With Placebo in Antidepressant Clinical Trials." *Archives of General Psychiatry* 57 (4): 311.

Kingma, E. 2010. "Paracetamol, Poison, and Polio: Why Boorse's Account of Function Fails to Distinguish Health and Disease." *The British Journal for the Philosophy of Science* 61 (2): 241–64.

Kirsch, I., B. J. Deacon, T. B. Huedo-Medina, A. Scoboria, T. J. Moore, and B. T. Johnson. 2008. "Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the

Food and Drug Administration." *PLoS Medicine* 5 (2): e45.

Kirsch, I., T.J. Moore, A. Scoboria, and S.S. Nicholls. 2002. "The Emperor's New Drugs: An Analysis of Antidepressant Medication Data Submitted to the US Food and Drug Administration." *Prevention and Treatment* 5 (1): n/a.

Kitcher, P. 2012. *Preludes to Pragmatism: Toward a Reconstruction of Philosophy*. New York and Oxford: Oxford University Press.

Kline, P. 1998. *The New Psychometrics: Science, Psychology and Measurement*. London: Routledge.

Kramer. 2016. *Ordinarily Well: The Case for Antidepressants*. New York, NY: Farrar, Straus and Giroux.

Krantz, D., R. D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of Measurement, Vol. I: Additive and Polynomial Representations*. San Diego and London: Academic Press.

Kristoffersen, I. 2010. "The Metrics of Subjective Wellbeing: Cardinality, Neutrality and Additivity." *Economic Record* 86 (272): 98–123.

Kuhn, T. 1974. "Second Thoughts on Paradigms." In *The Structure of Scientific Theories*, edited by F. Suppe. Urbana, IL: University of Illinois Press.

Kyngdon, A. 2008. "The Rasch Model from the Perspective of the Representational Theory of Measurement." *Theory & Psychology* 18 (1): 89–109.

———. 2011. "Plausible Measurement Analogies to Some Psychometric Models of Test Performance." *British Journal of Mathematical and Statistical Psychology* 64: 478–497.

Lacasse, J. R, and J. Leo. 2005. "Serotonin and Depression: A Disconnect between the Advertisements and the Scientific Literature." *PLoS Medicine* 2 (12): e392.

Lakens, D. 2013. "Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs." *Frontiers in Psychology* 4 (November): 1–12.

Lance, C. E., M. M. Butts, and L. C. Michels. 2006. "The Sources of Four Commonly Reported Cutoff Criteria What Did They Really Say?" *Organizational Research Methods* 9 (2): 202–20.

Laurence, S., and E. Margolis. 2003. "Concepts and Conceptual Analysis." *Philosophy and Phenomenological Research* 67 (2): 253–82.

Leahey, T. H. 1980. "The Myth of Operationism." *The Journal of Mind and Behavior* 1 (2): 127–43.

Leucht, S., H. Fennema, R. Engel, M. Kaspers–Janssen, P. Lepping, and A. Szegedi. 2013. “What Does the HAMD Mean?” *Journal of Affective Disorders* 148 (2–3): 243–48.

Levine, M. V. 1970. “Transformations That Render Curves Parallel.” *Journal of Mathematical Psychology* 7 (3): 410–43.

Loevinger, J. 1947. *A Systematic Approach to the Construction and Evaluation of Tests of Ability*. Edited by John F. Dashiell. *Psychological Monographs*. Vol. 61. Washington: The American Psychological Association.

———. 1957. “Objective Tests as Instruments of Psychological Theory.” *Psychological Reports* 3 (9): 635–94.

Lord, F. M., and M. R. Novick. 1968. *Statistical Theories of Mental Test Scores*. London and Reading, Mass.: Addison-Wesley.

Lovett, B.J., and B. Hood. 2011. “Realism and Operationism in Psychiatric Diagnosis.” *Philosophical Psychology* 24 (2): 207–22.

Luce, R. D. 1997a. “Several Unresolved Conceptual Problems of Mathematical Psychology.” *Journal of Mathematical Psychology* 41 (1): 79–87.

———. 1997b. “Quantification and Symmetry: Commentary on Michell, Quantitative Science and the Definition of Measurement in Psychology.” *British Journal of Psychology* 88 (3): 395–98.

Luce, R. D., D. Krantz, P. Suppes, and A. Tversky. 1990. *Foundations of Measurement, Vol. III: Representation, Axiomatization, and Invariance*. London and San Diego: Academic Press.

Luce, R. D., and L. Narens. 1987. “Measurement Scales on the Continuum.” *Science* 236 (19): 1527–32.

Luce, R. D., and J. W. Tukey. 1964. “Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement.” *Journal of Mathematical Psychology* 1 (1): 1–27.

Machamer, P., L. Darden, and C. F. Craver. 2000. “Thinking about Mechanisms.” *Philosophy of Science* 67 (1): 1–25.

Machery, E. 2009. *Doing without Concepts*. Oxford: Oxford University Press.

Machery, E., R. Mallon, S. Nichols, and S. P. Stich. 2004. “Semantics, Cross-Cultural Style.” *Cognition* 92 (3): B1–12.

Margolis, E., and S. Laurence. 1999. *Concepts: Core Readings*. Cambridge, Mass.: MIT Press.

———. 2007. “The Ontology of Concepts—Abstract Objects or Mental Representations?”

*Noûs* 41 (4): 561–93.

Mari, L. 2005. “The Problem of Foundations of Measurement.” *Measurement* 38 (4): 259–66.

Mari, L., P. Carbone, A. Giordani, and D. Petri. 2017. “A Structural Interpretation of Measurement and Some Related Epistemological Issues.” *Studies in History and Philosophy of Science Part A* 65–66 (October): 46–56.

Markus, K., and D. Borsboom. 2013. *Frontiers of Test Validity Theory : Measurement, Causation, and Meaning*. New York: Routledge.

Marmodoro, A., and D. Yates. 2016. *The Metaphysics of Relations*. Oxford: Oxford University Press.

Masters, G. N. 1988. “Item Discrimination: When More Is Worse.” *Journal of Educational Measurement* 25 (1): 15–29.

Matz, S. C., M. Kosinski, G. Nave, and D. J. Stillwell. 2017. “Psychological Targeting as an Effective Approach to Digital Mass Persuasion.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (48): 12714–19.

Maul, A., and J. McGrane. 2017. “As Pragmatic as Theft Over Honest Toil: Disentangling Pragmatism From Operationalism.” *Measurement: Interdisciplinary Research and Perspectives* 15 (1): 2–4.

Maul, A., D. Torres Irribarra, and M. Wilson. 2016. “On the Philosophical Foundations of Psychological Measurement.” *Measurement* 79: 311–20.

Maxwell, S. E., and H.D. Delaney. 1985. “Measurement and Statistics: An Examination of Construct Validity.” *Psychological Bulletin* 97 (1): 85–93.

McClimans, L. 2013. “The Role of Measurement in Establishing Evidence.” *Journal of Medicine and Philosophy* 38 (5): 520–38.

McClimans, L., J. Browne, and S. Cano. 2017. “Clinical Outcome Measurement: Models, Theory, Psychometrics and Practice.” *Studies in History and Philosophy of Science, Part A* 65–66: 67–73.

McLendon, H. J. 1955. “Uses of Similarity of Structure in Contemporary Philosophy.” *Mind* 64: 79–95.

Messick, S. 1995. “Validity of Psychological Assessment: Validation of Inferences from Persons’ Responses and Performances as Scientific Inquiry into Score Meaning.” *American Psychologist* 50 (9): 741–49.

Michell, J. 1986. “Measurement Scales and Statistics: A Clash of Paradigms.” *Psychological Bulletin* 100 (3): 398–407.

———. 1990. *An Introduction to the Logic of Psychological Measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

———. 2005. “The Logic of Measurement: A Realist Overview.” *Measurement* 38 (4): 285–94.

———. 2008. “Is Psychometrics Pathological Science?” *Measurement: Interdisciplinary Research and Perspectives* 6 (1–2): 7–24.

———. 2012. “Alfred Binet and the Concept of Heterogeneous Orders.” *Frontiers in Psychology* 3: 1–8.

Morgan, M., and M. Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

Moyal, A. M. 2001. *Platypus: The Extraordinary Story of How a Curious Creature Baffled the World*. Crows Nest N.S.W.: Allen & Unwin.

Mulaik, S. A. 1972. *Foundations of Factor Analysis*. New York and London: McGraw-Hill.

Narayan, D., R. Chambers, M.K. Shah, and P. Petesch. 2000. *Voices of the Poor - Crying Out for Change*. The World Bank.

Narens, L., and R. D. Luce. 1993. “Further Comments on the ‘Nonrevolution’ Arising from Axiomatic Measurement Theory.” *Psychological Science* 4 (2): 127–30.

Nersessian, N. J. 2008. *Creating Scientific Concepts*. London and Cambridge, Mass.: MIT Press.

Noury, J. Le, J.M. Nardo, D. Healy, J. Jureidini, M. Raven, C. Tufanaru, and E. Abi-Jaoude. 2015. “Restoring Study 329: Efficacy and Harms of Paroxetine and Imipramine in Treatment of Major Depression in Adolescence.” *BMJ (Clinical Research Ed.)* 351 (September): h4320.

Nozick, R. 1974. *Anarchy, State, and Utopia*. Oxford: Blackwell.

Nunnally, J. C., and I. H. Bernstein. 1994. *Psychometric Theory*. New York and London: McGraw-Hill.

Oishi, S. 2006. “The Concept of Life Satisfaction across Cultures: An IRT Analysis.” *Journal of Research in Personality* 40 (4): 411–23.

Orilia, F., and C. Swoyer. 2017. “Properties.” *Stanford Encyclopedia of Philosophy, (Winter 2017 Edition)*.

Osherson, D., and D. Lane. 2018. “Levels of Measurement.” Online Statistics: An Interactive Multimedia Course of Study. 2018.

Pavot, W., and E. Diener. 1993. “Review of the Satisfaction With Life Scale.” *Psychological Assessment* 5 (2): 164–72.

Perline, R., B. D. Wright, and H. Wainer. 1979. “The Rasch Model as Additive Conjoint Measurement.” *Applied Psychological Measurement* 3 (2): 237–55.

Pittenger, D. 2005. “Cautionary Comments Regarding the Myers-Briggs Type Indicator.” *Consulting Psychology Journal: Practice and Research* 57 (3): 210–21.

Putnam, H. 1970. “Is Semantics Possible?” *Metaphilosophy* 1 (3): 187–201.

Rammstedt, B., and O.P. John. 2007. “Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German.” *Journal of Research in Personality* 41 (1): 203–12.

Reason, P., and H. Bradbury. 2008. *The SAGE Handbook of Action Research: Participative Inquiry and Practice*. London: SAGE.

Reiss, J. 2008. *Error in Economics: Towards a More Evidence–Based Methodology*. London: Routledge.

Revelle, W., and R. E. Zinbarg. 2009. “Coefficients Alpha, Beta, Omega, and the Glb: Comments on Sijtsma.” *Psychometrika* 74 (1): 145–54.

Reynolds, W. M., and K. A. Kobak. 1995. “Reliability and Validity of the Hamilton Depression Inventory: A Paper-and-Pencil Version of the Hamilton Depression Rating Scale Clinical Interview.” *Psychological Assessment* 7 (4): 472.

Rhemtulla, M., D. Borsboom, and R. Van Bork. 2017. “How to Measure Nothing.” *Measurement* 15 (2): 95–97.

Rice, C. M. 2013. “Defending the Objective List Theory of Well-Being.” *Ratio* 26 (2): 196–211.

Russo, F., and J. Williamson. 2007. “Interpreting Causality in the Health Sciences.” *International Studies in the Philosophy of Science* 21 (2): 157–70.

Rust, J., and S. Golombok. 2009. *Modern Psychometrics: The Science of Psychological Assessment*. London: Routledge.

Saal, F. E., R. G. Downey, and M. A. Lahey. 1980. “Rating the Ratings: Assessing the Psychometric Quality of Rating Data.” *Psychological Bulletin* 88 (2): 413–28.

Scharp, K. 2013. *Replacing Truth*. Oxford: Oxford University Press.

Searls, D. 2017. *The Inkblots: Hermann Rorschach, His Iconic Test, and the Power of Seeing*. Crown.

Shafer, A. B. 2006. “Meta-Analysis of the Factor Structures of Four Depression Questionnaires: Beck, CES-D, Hamilton, and Zung.” *Journal of Clinical Psychology* 62 (1): 123–46.

Shepherd, J., and J. Justus. 2015. “X-Phi and Carnapian Explication.” *Erkenntnis* 80 (2): 381–402.

Sijtsma, K. 2009. “On the Use, the Misuse, and the Very Limited Usefulness of Cronbach’s Alpha.” *Psychometrika* 74 (1): 107–20.

Sireci, S. G. 1998. “The Construct of Content Validity.” *Social Indicators Research* 453 (1): 83–117.

Snyder, A. G., M. A. Stanley, D. M. Novy, P. M. Averill, and J. G. Beck. 2000. “Measures of Depression in Older Adults with Generalized Anxiety Disorder: A Psychometric Evaluation.” *Depression and Anxiety* 11 (3): 114–20.

Stegenga, J. 2018. *Medical Nihilism*. Oxford: Oxford University Press.

Stevens, S. S. 1934. “The Volume and Intensity of Tones.” *The American Journal of Psychology* 46 (3): 397.

———. 1935. “The Operational Definition of Psychological Concepts.” *Psychological Review* 42 (6): 517–27.

———. 1946. “On the Theory of Scales of Measurement.” *Science (New York, N.Y.)* 103 (2684): 677–80.

———. 1951. “Mathematics, Measurement, and Psychophysics.” In *Handbook of Experimental Psychology.*, edited by S. S. Stevens, 1–49. Oxford: Wiley.

Strauss, M. E., and Gr. T. Smith. 2009. “Construct Validity: Advances in Theory and Methodology.” *Annual Review of Clinical Psychology* 5 (1): 1–25.

Streiner, D. L. 2003. “Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency.” *Journal of Personality Assessment* 80 (1): 99–103.

Suppes, P., D. Krantz, R. D. Luce, and A. Tversky. 1989. *Foundations of Measurement, Vol. II: Geometrical, Threshold, and Probabilistic Representations*. San Diego and London: Academic Press.

Suppes, P., and J. L. Zinnes. 1962. “Basic Measurement Theory.” In *Handbook of Mathematical Psychology*, edited by R. D. Luce, R.R. Bush, and E. Galanter. Oxford: Wiley.

Sutton, J. 2004. “Are Concepts Mental Representations or Abstracta?” *Philosophy and Phenomenological Research* 68 (1): 89–108.

Tal, E. 2012. “The Epistemology of Measurement: A Model-Based Account.” University of Toronto.

———. 2016. “Making Time: A Study in the Epistemology of Measurement.” *The British Journal for the Philosophy of Science* 67 (1): 297–335.

———. 2017. “Measurement in Science.” *The Stanford Encyclopedia of Philosophy, (Fall 2017 Edition)*.

Tarescavage, A. M., J. Scheman, and Y. S. Ben-Porath. 2015. “Reliability and Validity of the Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF) in Evaluations of Chronic Low Back Pain Patients.” *Psychological Assessment* 27 (2): 433–46.

Thagard, P. 1999. *How Scientists Explain Disease*. Princeton, NJ: Princeton University Press.

Thomson, G. H. 1940. “The Nature And Measurement Of The Intellect.” *Teachers College Record*.

Tolman, E. C. 1936. “An Operational Analysis of Demands.” *Erkenntnis* 6: 383–92.

Traub, R. E. 2005. “Classical Test Theory in Historical Perspective.” *Educational Measurement: Issues and Practice* 16 (4): 8–14.

Tryon, R. C. 1957. “Reliability and Behavior Domain Validity: Reformulation and Historical Critique.” *Psychological Bulletin* 54 (3): 229–49.

Veenhoven, R. 2009. “The International Scale Interval Study: Improving the Comparability of Responses to Survey Questions About Happiness.” In *Quality of Life and the Millennium Challenge*, edited by V. Møller and D. Huschka, 45–58. Dordrecht: Springer Netherlands.

Velleman, P. F., and L. Wilkinson. 1993. “Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading.” *The American Statistician* 47 (1): 65–72.

Vessonen, E. 2017. “Psychometrics versus Representational Theory of Measurement.” *Philosophy of the Social Sciences* 47 (4–5): 330–50.

———. 2018. “The Complementarity of Psychometrics and the Representational Theory of Measurement.” *The British Journal for the Philosophy of Science* n/a (n/a): n/a.

Vet, H. C. W. de, C. B. Terwee, L. B. Mokkink, and D. L. Knol. 2011. *Measurement in Medicine*. Cambridge: Cambridge University Press.

Walsh, J. A. 1968. “Book Review of Statistical Theories of Mental Test Scores.” *Educational and Psychological Measurement* 28 (4): 1266–69.

Watson, J. B. 1913. “Psychology as the Behaviorist Views It.” *Psychological Review* 20 (2): 158–77.

Weinberg, J. M., S. Nichols, and S. Stich. 2001. “Normativity and Epistemic Intuitions.” *Philosophical Topics* 29 (1): 429–60.

Weinberg, J.M., C. Gonnerman, C. Buckner, and J. Alexander. 2010. “Are Philosophers Expert Intuiters?” *Philosophical Psychology* 23 (3): 331–55.

Wittgenstein, L. 2009. *Philosophical Investigations*. Translated by G. E. M. Anscombe. Edited by P. M. S. Hacker and J. Schulte. Chichester, West Sussex: John Wiley & Sons, Inc.

Woolgar, S. 1988. *Science: The Very Idea*. Chichester: Ellis Horwood.

Wright, B.D., and M. H. Stone. 1999. *Measurement Essentials*. Wilmington: Wide Range Inc.

Youyou, W., M. Kosinski, and D. Stillwell. 2015. “Computer-Based Personality Judgments Are More Accurate than Those Made by Humans.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (4): 1036–40.